The international journal of science / 10 September 2020

nature

GALACTIC CORONAE

Simulation reveals the origin of the Magellanic Stream

Coronavirus
How are viral mutations likely to affect the pandemic?

Genetic fairness
Data from minority populations should be used not discarded

Careers survey
Snapshot of postdoc life highlights the impact of COVID-19

Vol. 585, No. 7824
nature.com

The international journal of science / 10 September 2020

nature

Japan after Abe: time for a fresh start

Departing Prime Minister Shinzō Abe's successor needs to embrace diversity, diplomacy and better regulation in science.

When Shinzō Abe returned as Japan's prime minister in 2012, the country had been through five leaders in as many years, and had one of the developed world's more sluggish economies. Eight years later, as Japan's longest-serving leader steps down for health reasons, the country is more stable politically. But the changes Abe made in the name of economic growth and social development leave a mixed legacy. Despite a deliberate effort to boost the economy through science — particularly biomedical research — the growth rate see-sawed during the Abe years, and never exceeded its 2017 high of 2.3%.

On taking office, Abe, who leads the right-wing Liberal Democrat Party, vowed to get more out of science. Overall, more of Japan's researchers are now publishing with co-authors elsewhere, but the country's share of international science publishing has been declining for some years.

Japan spends 3.2% of its national income on research and development — one of the highest amounts among the G20 group of the world's largest economies (the United States spends 2.8%). But about 80% of this spending comes from industry. Japan's share of government investment in science remains low by developed-country standards. At the same time, Abe's efforts to squeeze more innovation — one of the components of economic growth — out of scientific research have produced no clear successes.

In 2015, Abe launched the Japan Agency for Medical Research and Development (AMED) — the nation's equivalent of the US National Institutes of Health, but with more focus on transferring discoveries to the clinic. In 2018, the agency's annual budget was ¥126.6 billion (US$1.2 billion).

Although it is too early to judge AMED's performance, the government had already moved to commercialize regenerative medicine. Two laws passed in 2014 allow companies to obtain faster regulatory approval to use stem cells and other regenerative therapies in patients. In permitting this, Japan decided to disregard the consensus of international experts that stem-cell treatments should not be commercialized until rigorous and unambiguous evidence — in the form of controlled clinical trials — confirms that they are safe and effective. Despite much criticism at home and abroad, Japan's government has not altered its approach.

Such assertive techno-nationalism is nothing new, and, if anything, seems to be becoming more common around the world. But this isn't how Japan has traditionally done things. Researchers have been resolute in their view that science must only ever be used for peaceful purposes and for economic development. But Abe sought to revisit that. After coming to power, he increased defence spending and tried — but ultimately failed — to amend Japan's pacifist constitution. He also launched a fund to support technologies with potential military uses, to be overseen by the defence ministry's Acquisition, Technology and Logistics Agency.

> "One of Abe's most notable failures has been in his government's inability to fulfil a promise to improve gender diversity in Japan's workplaces."

Japan's government has also been considering putting restrictions on international collaborations in areas such as quantum computing, artificial intelligence and semiconductor design. This would stop what it regards as sensitive scientific research from being shared with researchers in other countries, especially China, and is in line with policies being enacted in the United States and Australia.

Such actions, if carried out without due care and attention to the consequences, risk overturning progress towards Japan's long-held ambition to internationalize its research community. A significant share of the post-doctoral researchers and graduate students in Japan come from China, as do 40% of Japan's international exchange students. It would be a shame if science, which has played a small part in building bridges between the peoples of these two nations, were to drive a wedge between them.

Gender imbalance

One of Abe's most notable failures has been in his government's inability to fulfil a promise to improve gender diversity in Japan's workplaces. The five-year basic science and technology plan that started in 2016 embraced a national goal for women to make up 30% of the scientific workforce by 2020. As of 2019, only 16.6% of scientists were women, according to the Ministry of Internal Affairs and Communications. And this number remains among the lowest of the G20 countries — women constitute 28% of scientists in Germany, 39.5% in Russia and 45% in South Africa.

So what of the future? Japanese politics runs on consensus, and politicians — regardless of party affiliation — are not known for the kind of impulsive alpha-male leadership we are currently seeing in other countries. Consensus is an important and necessary characteristic in a political system, but it also means that when governments want or need to change course, this takes longer. And that means that, although Japan will soon get a new prime minister, the incoming administration is unlikely to immediately deviate strongly from the path set out by Abe.

In the long run, that is not the path Japan needs to follow. The country's researchers must persuade the incoming administration that the research system will become more innovative and resilient, not by fast-tracking technology regulation, but by embracing diversity and inclusion, a smarter approach to government investments, and better science diplomacy.

We are living through one of the most worrying and unpredictable periods in recent history, with an ever-present threat of conflict and tension between states. Japan has so far been a beacon to the world in its embrace of science for peace, and the world needs this remarkable country to stay that way.

nature

Postdocs in crisis: science risks losing the next generation

The pandemic has worsened the plight of postdoctoral researchers. Funders need to be offering more than moral support.

Postdoctoral researchers know what it's like to be in career limbo, spending years — in some cases decades — on a succession of short-term contracts. The anxiety and uncertainty this creates can be immense. And, as the results of a new *Nature* survey show, the pandemic is adding to postdocs' distress. The current generation might be facing the most severe career and health crisis so far.

Nature asked postdocs how the pandemic is affecting their current and future career plans; about their health and well-being; and whether they feel supported by their supervisors.

The poll ran in June and July, and more than 7,600 people responded from across 19 disciplines. The sample, a self-selecting group scattered over 93 countries, is not fully representative globally, because the overwhelming majority of respondents are in Europe and North America. But the picture that emerges is undoubtedly concerning.

Six out of ten respondents think the pandemic has worsened their career prospects, and one in four feel that their supervisors have not done enough to support them during the pandemic. Moreover, 23% of respondents said that they have sought help for anxiety or depression caused by their work, and a further 26% would like such help but have not yet sought it. This is in line with other findings of pandemic-related mental ill-health.

Equally concerning is the fact that 51% of respondents to the latest survey have considered leaving active research because of work-related mental-health concerns. It is tragic that so many early-career researchers are in such distress. And it spells trouble for knowledge, discovery and invention if so many people are concluding that they have no future in science.

The written survey responses offer a more detailed picture. An engineer in India wrote that he is unable to take up a postdoctoral job offer abroad because of travel restrictions imposed as a result of the pandemic; a researcher in Germany described how employment offers were being withdrawn; a physicist in Brazil feared that the government would curtail scholarships. These individual stories reflect the fact that universities, which are under financial pressure because of the pandemic, are widely freezing recruitment and cutting roles.

We put the survey findings to several major funding organizations in Australia, Europe and the United States, and asked what they are doing to help. They described the ways in which they are supporting early-career researchers, such as by providing extensions to project deadlines. With most national economies in recession, all efforts to help workers, no matter how small, are welcome; but, on their own, small measures will sadly not be enough to save many academic science careers.

> **"Now is the time to pause or slow down the treadmill of research evaluation."**

The US National Science Foundation (NSF), for example, said that it has extended project deadlines and directed universities to continue paying the salaries of NSF-funded postdocs while research has had to be put on hold. But it isn't clear who is funding these salary extensions. The NSF isn't providing any extra money, and universities are not compelled to comply with the NSF's requests — nor should they be. Other funders provided a similar response to our questions: grants are being extended, but there is no more money from the funder. This is neither fair nor sustainable.

Universities cannot be expected to bear this extra cost. The pandemic is already severely testing their finances, especially for those institutions that rely on income from international students' fees. Global student mobility will be much lower than usual in the coming academic year, and some institutions will lose a good fraction of their fee income as a result. In places where research is cross-subsidized from tuition-fee income, contract-research workers such as postdocs are most vulnerable to losing their jobs — and, in many fields, that will disproportionately affect women and people from minority groups, who constitute a comparatively high share of the postdoctoral workforce.

Such uncertainty is adding to the strain being experienced by postdocs, who rightly worry that shuttered experiments and unfinished manuscripts will set back their quest for grants and jobs. And our poll results suggest that many are looking to leave their posts now, anticipating that worse is to come. Research and university leaders must think of innovative ways to support early-career colleagues.

Senior investigators who wish to see promising younger colleagues find long-term careers in academia must look for ways to make it possible for them to stay. But they must equally be champions for those who want to pursue fulfilling careers in science elsewhere. What matters is that talented people find satisfying careers in science. Principal investigators should show flexibility, patience and support for everyone in their group. They and their institutions must also push harder than ever for accessible mental-health services.

Now is also the time to pause or slow down the treadmill of research evaluation. Even before the pandemic, early-career researchers faced the pressures of continuous assessment, and a more competitive and less secure working atmosphere than those who came before them. The pandemic has worsened this situation. A crushing, urgent crisis for individuals now risks becoming an existential crisis for a system that needs today's postdocs to become tomorrow's research leaders in academia, industry, government and the non-profit world. We cannot allow the pandemic to destroy the careers of these smart young people — many of whom are likely to contribute to finding a solution to it.

World view

By Joshua Sharfstein

How the FDA should protect its integrity from politics

Long before the pandemic, the agency set criteria to ensure science drives its decisions.

On my first day as acting commissioner of the US Food and Drug Administration (FDA), in March 2009, I walked into an agency that was under a cloud. That month, news reports had alleged that the FDA had stumbled in its crucial role of protecting the US public from unsafe treatments, because it had cleared a medical device for use in patients after "a lobbying campaign that overcame repeated rejections by scientists". I decided to investigate whether the agency's integrity had indeed been compromised by politics. Doing so meant first answering this question: what is meant by integrity at the FDA?

Today, the answer is more crucial than ever. In April this year, a whistle-blower lawsuit was filed alleging inappropriate pressure from the White House to promote an unproven treatment (the malaria drug hydroxychloroquine) for COVID-19. Under pressure from President Donald Trump, the FDA had issued an emergency-use authorization for the medication in March, only to backtrack weeks later.

In August, the FDA commissioner, Stephen Hahn, stood with the president on the eve of the Republican National Convention to announce the authorization of convalescent plasma as treatment for COVID-19. The president made misleading statements about the evidence supporting this treatment and asserted without evidence that FDA staff were holding up approvals for political reasons.

To define integrity at the FDA a decade ago, I turned to the agency's chief scientist, top lawyer and leading policy official. They set out three criteria (see go.nature.com/2gx1hz). The first was that decisions should be "based on a rigorous evaluation of the best available science", drawing on "appropriate expertise, including the use of advisory committees". Today, the agency has yet to consult such a committee for a major decision on COVID-19. Instead, criticism of FDA actions from non-agency scientists, including the leaders of the US National Institutes of Health, has filtered into news reports, sowing doubts about whether potential risks and unintended consequences have been properly considered.

The second criterion was that decisions should be "reached and documented through a process that promotes open-mindedness", with the "bases of final decisions and processes for decision-making … adequately documented and explained". In other words, transparency is crucial to integrity; without seeing the evidence and hearing the reasoning, people often assume the worst.

Globally, the lack of transparency about decision-making is eroding trust in many governments whose response to the pandemic has been poor. The FDA has disclosed little about how it is making decisions, squandering the chance to build up understanding and support. During my time at the FDA, agency leaders met challenges, such as debates about the safety of diabetes medicines, by releasing detailed memos, publishing explanatory articles in medical journals and giving press interviews.

> "Will the voice of the FDA just be one of many tainted by politics?"

The third criterion of integrity was that decisions should be "made without inappropriate or external interference". It stipulated that "data and opinions are not suppressed, distorted, or manipulated" and that "pressure from external persons does not influence the regulatory decision".

There can be no doubt that Trump's attacks aim to influence decision-making at the agency. Last month, he alleged, without evidence, that "the deep state, or whoever, over at the FDA" is stalling interventions for COVID-19. His chief of staff has publicly stated that the president wants the agency to "feel the heat".

Back in September 2009, the FDA released preliminary results of its investigation into the clearance of the controversial device (see go.nature.com/3jan9nj). The report detailed multiple departures from "processes, procedures, and practices" during review, the exclusion of key staff from the scientific debate, and a "failure to respond appropriately to external pressure on decision-makers". The agency, led by Margaret Hamburg (I was then principal deputy commissioner), took the conclusions seriously. We moved to revoke the clearance, pledging to close identified gaps and improve the review process for all medical devices. Charles Grassley, Republican senator for Iowa and a frequent FDA critic, stated: "The kind of reflection and the commitment to action made in this report is key to the FDA building public confidence."

How can the FDA defend its integrity today? One positive sign is Hahn's commitment to hold an advisory-committee meeting before approving or authorizing a vaccine for COVID-19. Hahn has also stated on several occasions that the agency will make decisions only on the basis of "good science and sound data". Beyond anodyne assurances, however, Hahn should reject political pressure from the White House; set out in detail the process for vaccine review; and commit to releasing key data and decision memos.

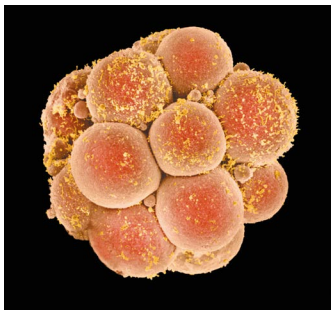
Integrity is central to the FDA's credibility. Patients and clinicians treating COVID-19 are already making judgement calls with limited evidence. Soon, amid a cacophony of misinformation and confusion, amplified by political polarization and social media, Americans will have to weigh the merits of vaccination. Will the voice of the FDA just be one of many tainted by politics? Or can it provide the clarity that the moment demands?

With the number of US deaths from COVID-19 approaching 200,000, the integrity of the country's leading public-health regulatory agency is more than an abstraction; it is a matter of life and death.

Joshua Sharfstein is vice-dean for public health practice and community engagement at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. e-mail: joshua.sharfstein@jhu.edu

JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH

News in brief



'CRISPR BABIES' STILL TOO RISKY, SCIENTISTS WARN

The use of gene editing in human embryos could one day prevent some serious genetic disorders from being passed from parents to their children. But, for now, the technique is too risky to be used in embryos destined for implantation, says a high-profile international commission. And even when the technology is mature, the commission says, its use should initially be permitted in only a narrow set of circumstances.

The recommendations, released on 3 September, come from a panel of experts convened by the US National Academy of Medicine, the US National Academy of Sciences and the UK Royal Society. The panel's report reviewed the scientific and technical state of heritable gene editing, rather than ethical questions. It joins a wealth of reports that have argued against using gene editing in the clinic until researchers can address safety worries.

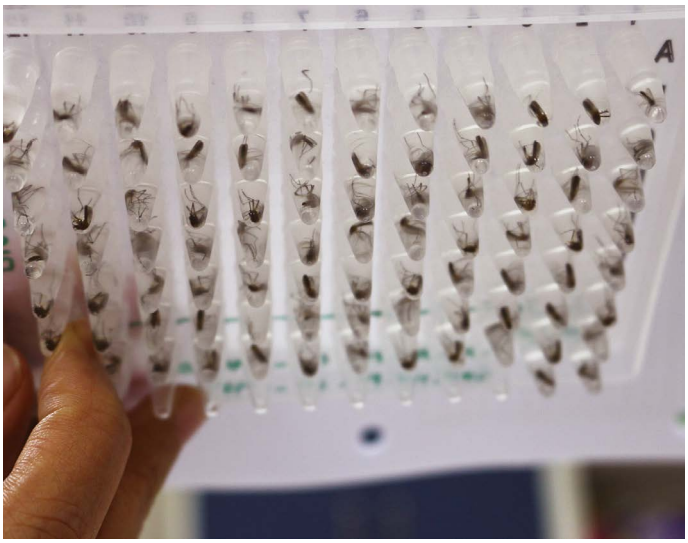
The commission was formed after Chinese biophysicist He Jiankui shocked the world in 2018 by announcing that he had edited human embryos that were then implanted in women's wombs, in an effort to make the resulting children resistant to HIV. The work led to the birth of two children with edited genomes, but was condemned by scientists. He and two of his colleagues received prison sentences.

STEROID DRUGS LINKED TO LOWER COVID-19 MORTALITY

People severely ill with COVID-19 are less likely to die if they are given drugs called corticosteroids than are people who are not, according to an analysis of hospital patients on five continents.

Earlier findings showed that the steroid dexamethasone cut deaths in people with COVID-19 on ventilators. To examine the effects of steroids in general, Jonathan Sterne at the University of Bristol, UK, and his colleagues did a meta-analysis that pooled data from seven clinical trials. Each of the seven studied the use of steroids in people who were critically ill with COVID-19 (REACT Working Group *J. Am. Med. Assoc.* https://doi.org/d7z8; 2020). The trials included more than 1,700 people across 12 countries.

The team analysed participants' status 28 days after they were randomly assigned to take either a steroid or a placebo. The risk of death was 32% for those who took a steroid and 40% for those who took a placebo. The authors say that steroids should be part of the standard treatment for people with severe COVID-19.



MODIFYING MOSQUITOES COULD ELIMINATE DENGUE

Cases of dengue fever plummeted by more than 75% in the Indonesian city of Yogyakarta after researchers released mosquitoes modified to carry *Wolbachia* bacteria, which stop the insects from transmitting some viruses. These results are the strongest evidence yet that the *Wolbachia* technique, in development since the 1990s, could rid the world of some deadly mosquito-borne diseases, researchers say.

Wolbachia-carrying mosquitoes were released over a six-month period in randomly designated parts of Yogyakarta, starting in 2016. Rates of dengue in these areas were 77% lower, as assessed during several years after release, compared with areas that did not receive the mosquitoes. The results were reported in press releases on 26 August, but the full data underlying the figures are yet to be published.

It will be important to scrutinize the full data, but "a 77% reduction is really extraordinary", says Philip McCall, a vector biologist at the Liverpool School of Tropical Medicine, UK. "This does have huge promise."

Around 60% of insect species carry *Wolbachia pipientis*, but the bacterium does not naturally infect the *Aedes aegypti* mosquito species that transmits dengue, Zika and numerous other viruses. Beginning in the 1990s, researchers developed laboratory populations of *Wolbachia*-infected *A. aegypti* and showed that these insects do not transmit viruses, including dengue.

The Yogyakarta trial was coordinated by the non-profit World Mosquito Program, which hopes to release *Wolbachia*-carrying mosquitoes in areas covering 75 million people at risk of dengue in the next 5 years, and to reach half a billion people in a decade.

L TO R: YORGOS NIKAS/SPL; WORLD MOSQUITO PROGRAM

News in focus

SATISH BATE/*HINDUSTAN TIMES*/GETTY

A vaccine will be key to controlling India's coronavirus outbreak.

INDIA'S ROLE AS GLOBAL COVID-19 VACCINE MAKER MAY NOT HELP ITS PEOPLE

The country will struggle to manufacture and distribute enough doses to control its own massive outbreak, scientists say.

By Gayathri Vaidyanathan

As scientists edge closer to creating a vaccine against the coronavirus SARS-CoV-2, Indian pharmaceutical companies are front and centre in the race to supply the world with an effective product. But researchers worry that, even with India's experience as a vaccine manufacturer, its companies will struggle to produce enough doses sufficiently fast to bring its own huge outbreak under control. On top of that, it will be an immense logistical challenge to distribute the doses to people in rural and remote regions.

Indian drug companies are major manufacturers of vaccines distributed worldwide, particularly those for low-income countries, providing more than 60% of vaccines supplied to the developing world. Because of this, the companies are likely to gain early access to any COVID-19 vaccine that works, says Sahil Deo, co-founder of India's CPC Analytics in Pune, which is studying vaccine distribution in the country.

Several Indian vaccine makers already have agreements to manufacture coronavirus immunizations that are being developed by international drug companies, or are working on their own vaccines. The government has said that these manufacturers can export some of their supplies as long as a proportion remains in the country.

Without India, there won't be enough vaccines to save the world, said Peter Piot, director of the London School of Hygiene and Tropical Medicine, during an online vaccine symposium organized by the Indian government in July.

A vaccine will be essential to combat India's huge coronavirus outbreak. On 6 September, the country reported more than 91,000 new cases — the highest single-day increase recorded in any country. By next year, the outbreak is predicted to be the world's largest.

To reduce the number of people dying from COVID-19, researchers say, those most at risk of exposure or severe infection will need to be immunized first. This includes first responders, people with other illnesses and older adults, who make up roughly 30% of the

population — around 400 million people, says Gagandeep Kang, a vaccinologist at the Christian Medical College in Vellore, India. But that is a huge number of vaccine doses that need to be made and shared out, researchers say.

The government has assembled a task force to determine how best to distribute the vaccines. It is headed by Vinod Paul, a member of the National Institution for Transforming India, a government think tank, and has representatives from state and central government agencies. The government is also working with vaccine makers to speed up clinical trials and regulatory approvals.

World's supplier

The world's largest vaccine maker, the Serum Institute of India in Pune, has an agreement to manufacture one billion doses of a coronavirus vaccine being developed by scientists at the University of Oxford, UK, and UK pharmaceutical company AstraZeneca if it is approved for use. The vaccine is currently undergoing phase III clinical trials in Brazil, the United Kingdom and the United States to test its effectiveness.

If the vaccine works, the Serum Institute and the Indian government have committed to reserve half the company's stock of it for India, and to supply half to low-income nations through Gavi, the Vaccine Alliance, which funds immunizations for low-income nations, says Adar Poonawalla, Serum's chief executive.

So far, the company has invested 11 billion rupees (US$200 million) to manufacture the vaccine, Poonawalla says, and it has produced about 2 million doses for use in regulatory clearances and testing, even before the trials have ended. Two factories that were producing other vaccines have been redirected to this effect, and the company can make 60 million to 70 million doses a month at full capacity, says Poonawalla.

The decision to stockpile the Oxford vaccine "has been solely taken to have a jump-start on manufacturing, to have enough doses available if the clinical trials prove successful", says Poonawalla. If the vaccine doesn't work, Serum will shift its attention to other candidates, he adds. The company is also developing and testing four other COVID-19 vaccines — two developed through in-house initiatives and two being developed in collaboration with biotechnology companies Novovax in Gaithersburg, Maryland, and Codagenix in Farmingdale, New York.

Drug firm Biologicals E, headquartered in Hyderabad, India, has also entered into a partnership to manufacture a vaccine candidate. This one is being developed by pharmaceutical company Janssen, a subsidiary of Johnson & Johnson based in Beerse, Belgium, and is currently going through early-stage safety trials. Biologicals E might also manufacture a candidate being developed by Baylor College of Medicine in Houston, Texas, the company says. And Indian Immunologicals, also in Hyderabad, is working with Australia's Griffith University in Brisbane to test and manufacture the university's vaccine. Two other Indian companies — Hyderabad-based Bharat Biotech and Zydus Cadila in Ahmedabad — are working on vaccines that are in phase I and II safety trials.

Scientists have applauded the Indian government for allowing the country's pharmaceutical companies to export some of their vaccine stocks to other nations. The decision to share supplies contrasts with the stance of countries such as the United States and the United Kingdom, which have each pre-ordered hundreds of millions of doses of coronavirus vaccines under development, enough to supply their respective populations many times over.

"The vaccine has to be kept cold, people have to be trained."

But even with manufacturers' commitment to supply a portion of their vaccines locally, scientists say that making the required 400 million doses for people who are most at risk of contracting severe COVID-19 will still take time. And by that point, the brunt of the epidemic, which is currently in major cities, will probably have shifted to rural areas, where health services are weaker, says Deo.

This means that the biggest hurdle will be getting vaccines to people across India. "It is a huge challenge," says Randeep Guleria, director of the All India Institute of Medical Sciences in New Delhi and a member of the government's vaccine task force. "India is a huge country, we have a very large population and we have remote areas, like the Northeast and Ladakh" in the Himalayas.

The immunization programme will probably take years, says Kang. One of the country's largest vaccination campaigns so far — delivery of the measles–rubella vaccine to 405 million children, starting in 2017 — has taken 3 years.

Guleria says that innovative approaches will be needed to distribute vaccines in rural and remote regions. He says national election campaigns could offer lessons. In 2019, 11 million poll workers journeyed across India to set up polling stations, so that people didn't need to travel more than 2 kilometres to vote. The network reached 900 million voters, including those in the most remote areas, in just over 6 weeks. A similar network of health officials to give vaccines could cover much of the country, says Guleria.

But it's not as simple as getting the vaccine to people, says Kang. "The vaccine has to be kept cold, people have to be trained." It will also be expensive to buy syringes and needles, to train people to vaccinate, and to purchase the vaccine.

The Serum Institute has priced the Oxford vaccine at 225 rupees (US$3) a dose. That means the cost of vaccinating 400 million people will be at least $1.2 billion. Typically, the government buys vaccines for less than the price of bottled water — 60 rupees. It's unlikely that the Indian government will bear the entire cost of immunizing its people, Deo notes. It will probably pay for vaccinations for the poorest citizens, and ask everyone else to buy their own vaccines, he says.

COVID-19 REINFECTION: THREE QUESTIONS SCIENTISTS ARE ASKING

Repeat infections raise questions about long-term immunity and the prospects for a vaccine.

By Heidi Ledford

When news broke last month that a man living in Hong Kong had been infected with the coronavirus again, months after recovering from a previous bout of COVID-19, immunologist Akiko Iwasaki had an unusual reaction. "I was really kind of happy," she says. "It's a nice textbook example of how the immune response should work."

For Iwasaki, who has been studying immune responses to the SARS-CoV-2 virus at Yale University in New Haven, Connecticut, the case was encouraging because the second infection did not cause symptoms. This, she says, suggested that the man's immune system might have remembered its previous encounter with the virus and fought off the repeat infection before it could do much damage.

But less than a week later, her mood shifted. Public-health workers in Nevada reported another reinfection — this time with more severe symptoms. Was it possible that the

immune system had not only failed to protect against the virus, but had also made things worse? "The Nevada case did not make me happy," Iwasaki says.

Duelling anecdotes are common in the see-saw world of the COVID-19 pandemic, and Iwasaki knows that she cannot draw firm conclusions about long-term immune responses to SARS-CoV-2 from just a few cases. But in the coming weeks and months, Iwasaki and others expect to see more reports of reinfection, and, in time, a more detailed picture could emerge.

As data trickle in, *Nature* runs through the key questions that researchers are trying to answer about reinfection.

How common is reinfection?

Reports of possible reinfections have circulated for months, but the recent findings are the first to seemingly rule out the possibility that a second infection was merely a continuation of a first.

To establish that in each person, the two infections were separate events, both the Hong Kong and Nevada teams sequenced the viral genomes from the first and second infections. Both found enough differences to convince them that separate variants of the virus were at work.

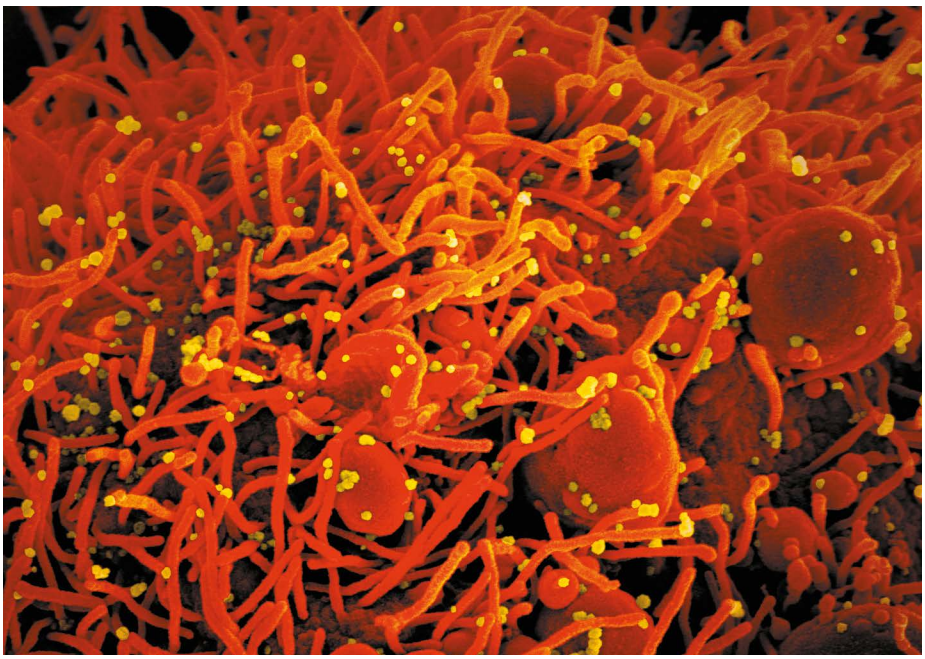
But, with only two examples, it is still unclear how frequently reinfections occur. And with 26 million known coronavirus infections worldwide so far, a few reinfections might not be cause to worry – yet, says virologist Thomas Geisbert at the University of Texas Medical Branch in Galveston. We need a lot more information on how prevalent this is, he says.

That information might be on the horizon: timing and resources are converging to make it possible to identify more instances of reinfection. Some regions are experiencing fresh outbreaks, providing an opportunity for people to be re-exposed to the virus. Testing has also become faster and more available.

And scientists in public-health laboratories are beginning to find their feet again, says Mark Pandori, director of the Nevada State Public Health Laboratory in Reno, and an investigator on the Nevada team. During the first wave of the pandemic, it was hard to imagine tracking reinfections when testing labs were overwhelmed. Since then, Pandori says that his lab has had time to breathe – and to set up sequencing facilities that can rapidly sequence large numbers of viral genomes from positive SARS-CoV-2 tests.

How severe are reinfections?

Unlike Iwasaki, virologist Jonathan Stoye at the Francis Crick Institute in London took no comfort from the lack of symptoms in the Hong Kong man's second infection. Drawing conclusions from a single case is hard, he says. "I'm not certain that really means anything at all." Stoye notes that the severity of COVID-19 varies enormously from person to person, and might also vary from infection to infection in the same person. Variables such as the initial dose of virus, possible differences between variants of SARS-CoV-2 and changes in a person's overall health could all affect the severity of a reinfection.



NIAID/NIH/SPL

Electron microscope image of SARS-CoV-2 coronavirus particles (yellow) on a cell (red).

Sorting out whether 'immunological memory' affects symptoms during a second infection is crucial, particularly for vaccine development. If symptoms are generally reduced the second time, that suggests the immune system is responding as it should.

"Reinfection shouldn't scare people. It shouldn't imply that a vaccine is not going to be developed."

But if symptoms are consistently worse during a second bout of COVID-19, the immune system might be making things worse, says immunologist Gabrielle Belz at the University of Queensland and the Walter and Eliza Hall Institute of Medical Research in Melbourne, Australia. For example, some cases of severe COVID-19 are worsened by rogue immune responses that damage healthy tissue. People who have experienced this during a first infection might have immune cells that are primed to respond in a disproportionate way again the second time, says Belz.

Another possibility is that antibodies produced in response to SARS-CoV-2 help, rather than fight, the virus during a second infection. This phenomenon, called antibody-dependent enhancement, is rare – but researchers found worrying signs of it while trying to develop vaccines against the coronaviruses responsible for severe acute respiratory syndrome and Middle East respiratory syndrome.

What does this mean for vaccines?

Historically, the vaccines that have been easiest to make are against diseases in which primary infection leads to lasting immunity, says Richard Malley, a paediatric infectious-disease specialist at Boston Children's Hospital in Massachusetts. Examples include measles and rubella.

But the capacity for reinfection does not mean that a vaccine against SARS-CoV-2 can't be effective, he adds. Some vaccines, for example, require 'booster' shots to maintain protection. "It shouldn't scare people," Malley says. "It shouldn't imply that a vaccine is not going to be developed or that natural immunity to this virus can't occur."

As public-health officials grapple with the dizzying logistics of vaccinating the world's population, a booster shot would hardly be welcome news, but it would not place long-term immunity against SARS-CoV-2 completely out of reach, says Malley. However, he is concerned about the possibility that vaccines will only reduce symptoms during a second infection, rather than prevent that infection altogether. This could effectively turn vaccinated people into asymptomatic carriers, putting vulnerable populations at risk.

For this reason, Malley is keen to see data on how much virus people 'shed' when reinfected with SARS-CoV-2. "They could still serve as an important reservoir of a future spread," he says. "We need to understand that better following natural infection and vaccination if we want to get out of this mess."

DAN KITWOOD, NICHOLAS KAMM/AFP/GETTY

Tensions between Presidents Xi Jinping and Donald Trump are spilling into research.

ARRESTS OF CHINESE SCIENTISTS MARK NEW FRONT IN US CRACKDOWN

US authorities increase scrutiny of visiting researchers' ties to Chinese military.

By Nidhi Subbaraman

When cancer researcher Juan Tang took refuge from the US Federal Bureau of Investigation (FBI) at the Chinese consulate in San Francisco in July, she drew national attention. Days later, the FBI arrested Tang — a Chinese national who was on a months-long research assignment in the United States — on charges of concealing her role as a Chinese military officer from the US government. Tang has since entered a not-guilty plea and is awaiting a jury trial.

Around the time of her arrest, the US authorities announced the arrests of a handful of other Chinese scientists for allegedly hiding ties they had to China's military on visa applications. Scholars of US–China policy say that the arrests mark a new front in the United States' battle against foreign interference in its universities, in which government officials are increasingly scrutinizing researchers' links to China's People's Liberation Army (PLA).

Scientists with ties to the Chinese military have visited the United States for years, says Brad Farnsworth, vice-president of the American Council on Education in Washington DC — but only now are officials "really looking very carefully at the background of the people who come here, particularly from China". Exactly how the FBI and the US Department of Justice (DoJ) are focusing their investigations remains unclear. The lack of concrete information from US authorities has triggered concerns that some scientists might be unfairly accused of espionage.

Many of the top hospitals in China, for example, are affiliated with the military, says Mary Gallagher, a political scientist at the University of Michigan in Ann Arbor, who studies US–China relations. "And so by default, if you're a doctor at one of those hospitals, you're going to have an affiliation with the Chinese military." That affiliation doesn't automatically mean that if you're collaborating with a US researcher you're engaging in espionage, she says.

"To assume a comprehensive conspiracy is too far from the reality."

The arrests come as tensions escalate between the United States and China. In 2018, US President Donald Trump's administration announced the China Initiative, aimed at stopping China from stealing intellectual property and technologies from US companies and research laboratories.

The US government's recent focus on researchers' links to the PLA has arisen alongside Chinese President Xi Jinping's 'military–civil fusion' strategy, in which university research and corporate intellectual property are being tapped for military use. In May, the Trump administration issued an order that would reject visa applications from researchers and students from some military-linked Chinese institutions, barring those people from entering the United States.

A new chapter

The arrests announced in July all involved accusations of visa fraud, according to officials at the DoJ and the FBI.

Tang had been a visiting researcher at the Department of Radiation Oncology at the University of California, Davis, since January. DoJ officials claim Tang denied serving in the military on her visa application — but that she is a "uniformed officer" in the PLA Air Force, a claim based in part on photographs of her in a military uniform that the DoJ submitted alongside the charges. The agency also claimed that the other researchers whose arrests were announced in July had past or current appointments in the Chinese military that they misrepresented on their visa applications.

The extent to which US research is actually being funnelled to the Chinese military, and how to block it meaningfully and fairly if it is, remain unclear, say experts — as do the parameters the United States is now using to label foreign scientists and collaborations as a threat.

According to court filings, one of the researchers arrested in July was working on military radar technology. But otherwise, the five scientists' fields of research alone — neurobiology, cell biology, medicine, physics, and machine learning — would not raise alarm from a national-security perspective, experts say.

Federal agents have not been transparent about what kind of US–China collaborations they view as risky. Glenn Tiffert, a research fellow at the conservative Hoover Institution, a public-policy think tank at Stanford University in California, suspects that there are many other cases that the government deems problematic from a national-security perspective.

To estimate the scope of the US government's concerns, Tiffert and his colleagues at Hoover released in July an analysis of Chinese- and English-language academic studies from 2013–19 that were listed in a major Chinese science and technology publishing database. The analysis found 254 that were co-authored by at least one scientist from a US university and one from a 'Seven Sons' university in China — seven institutions that were founded by or assisted the military before becoming civilian centres of higher education.

But an analysis by *Nature* using the Dimensions database from London-based Digital Science suggests that links between Chinese and US scientists are more prevalent than the Hoover report indicated. (Digital Science is part of Holtzbrinck, the majority shareholder in *Nature*'s publisher, Springer Nature). The analysis found more than 12,000 publications from 2015 to 2019 that had been co-authored by scientists in the United States and at one of the Seven Sons. Among those, 499 authors had a dual affiliation with a US institution and a Seven Sons university and were listed on papers declaring grant funding from the NIH or the US National Science Foundation.

But separating true threats from ordinary collaborations could be a challenge, some experts say. It has not been unusual for Chinese researchers with appointments in the military to visit the United States and work on non-classified projects, says Denis Simon, senior adviser to the president at Duke University in Durham, North Carolina. Simon led the Duke Kunshan University in China as vice-chancellor until July this year. "To assume a comprehensive conspiracy is too far from the reality," he says.

In general, universities do not have rules that bar scientists with affiliations to the foreign military from working with university researchers. But in the absence of nuanced federal guidelines, institutions might well be forced to take a fresh look at these collaborations.

"There is no longer any status quo to go back to," says Farnsworth.

Additional reporting by RIchard van Noorden.

ASTRONOMERS DETECT 'MINDBOGGLING' BLACK-HOLE COLLISION

Gravitational waves suggest merging black holes fell into 'forbidden' range of masses.

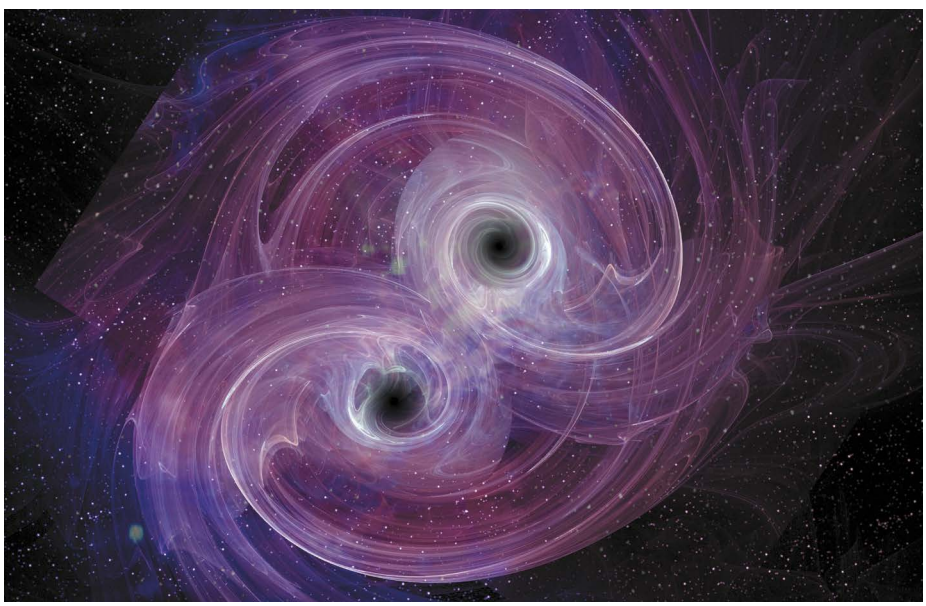
By Davide Castelvecchi

Astronomers have detected the most powerful, most distant and most perplexing collision of black holes yet, using gravitational waves. Of the two behemoths that fused when the Universe was half its current age, at least one — weighing 85 times as much as the Sun — has a mass that was thought to be too large to be involved in such an event. And the merger produced a black hole of nearly 150 solar masses, the researchers estimated, putting it in a range where no black holes had ever been conclusively seen before.

"Everything about this discovery is mindboggling," says Simon Portegies Zwart, a computational astrophysicist at Leiden University in the Netherlands. In particular, he says, the formation of the 150 solar mass black hole confirms the existence of 'intermediate mass' black holes: objects much more massive than a typical star, but not quite as big as the supermassive black holes that inhabit the centres of galaxies.

Ilya Mandel, a theoretical astrophysicist at Monash University in Melbourne, Australia, calls the finding "wonderfully unexpected".

The event, described in two papers published on 2 September[1,2], was detected on 21 May 2019, by the twin detectors of the Laser Interferometer Gravitational-Wave Observatory (LIGO) at Hartford, Washington, and Livingston, Louisiana, and by the smaller Virgo observatory near Pisa, Italy. It is named GW190521 after its detection date.



CAROL & MIKE WERNER/VISUALS UNLIMITED/SPL

An artist's impression of two colliding black holes.

Forbidden masses

Since 2015, LIGO and Virgo have provided new insights into the cosmos by sensing gravitational waves. These ripples in the fabric of space-time can reveal events such as the mergers of black holes that would not normally be visible with ordinary telescopes.

From the properties of the gravitational waves, such as how they change in pitch, astrophysicists can estimate the sizes and other features of the objects that produced them as these objects spiralled into each other. This ability has revolutionized the study of black holes, providing direct evidence for dozens of these objects, ranging in mass from a few to about 50 times the mass of the Sun.

These masses are consistent with black holes that formed in a 'conventional' way — when a very large star runs out of fuel to burn and collapses under its own weight. But the conventional theory says that stellar collapse should not produce black holes of about 65–120 solar masses. That's because towards the end of their lives, stars in a certain range of sizes become so hot at their centres that they start converting photons into pairs of particles and antiparticles — a phenomenon called pair instability. This triggers the explosive fusion of oxygen nuclei, which rips the star apart, completely disintegrating it.

In their latest discovery, the LIGO and Virgo detectors sensed only the last four ripples produced by the spiralling black holes, with a frequency that rose from 30 to 80 Hertz within one-tenth of a second. Whereas smaller black holes continue to 'chirp' up to higher frequencies, very large ones merge before this point, and barely enter the lower end of the frequency range to which the detectors are sensitive.

In this case, the two objects were estimated to weigh around 85 and 66 solar masses. "This is quite neatly in the range one would expect the pair-instability mass gap should be," says

LIGO astrophysicist Christopher Berry at Northwestern University in Evanston, Illinois.

Selma de Mink, an astrophysicist at Harvard University in Cambridge, Massachusetts, puts the cut-off for pair instability even lower, perhaps at 45 solar masses, which would push the lighter of the two objects firmly into the forbidden zone, too. "For me, both black holes are uncomfortably massive," she says.

Unconventional black holes

To explain their observations, the LIGO researchers considered a range of possibilities, including that the black holes had been around since the beginning of time. For decades, researchers have conjectured that such 'primordial' black holes could have spontaneously formed in a broad range of sizes shortly after the Big Bang.

"For me, both black holes are uncomfortably massive."

The main scenario the team contemplated is that the black holes got so large because they were themselves the result of earlier black-hole mergers. Black holes resulting from stellar collapse exist inside dense stellar clusters, and could undergo repeated mergers in principle. But this scenario is problematic because the black hole resulting from a first merger should typically get a kick from the gravitational waves and eject itself from the cluster. Only in rare cases would the black hole stay in an area where it could undergo another merger.

Successive mergers would be more likely if the black holes inhabited the crowded central region of their galaxy, de Mink says, where gravity is strong enough to prevent recoiling objects from shooting out.

It is not known in which galaxy the merger happened. But in roughly in the same region of the sky, a team of researchers spotted a quasar — an extremely bright galactic centre powered by a supermassive black hole — undergoing a flare about a month after the GW190521 signal[3]. The flare could have been a shockwave in the quasar's hot gas produced by the recoiling black hole, although many astronomers are cautious about accepting that the two phenomena are related.

This is the second time this year that the LIGO–Virgo collaboration has waded into a 'forbidden' mass range: in June, it described a merger involving an object of about 2.6 solar masses — typically considered too light to be a black hole but too massive to be a neutron star[4].

1. Abbott, R. *et al. Phys. Rev. Lett.* **125**, 101102 (2020).
2. Abbott, R. *et al. Astrophys. J.* **900**, L13 (2020).
3. Graham, M. J. *et al. Phys. Rev. Lett.* **124**, 251102 (2020)
4. Abbott, R. *et al. Astrophys. J.* **896**, L44 (2020).

Q&A

JACQUELINE SAUZIER

Cleaning up after Mauritius oil spill

The cargo ship MV *Wakashio* unleashed a vast oil spill when it ran aground on a coral reef on the southeast tip of Mauritius in the Indian Ocean in late July. The Japanese-owned vessel held 200 tonnes of diesel and 3,900 tonnes of fuel oil, an estimated 1,000 tonnes of which leaked into the sea when the ship's hull cracked on 6 August. It is the first reported spill of a new type of low-sulfur fuel that has been introduced to reduce air pollution. The spill has smeared oil over a 15-kilometre stretch of the coastline — an internationally recognized biodiversity hotspot. Jacqueline Sauzier, president of the non-profit Mauritius Marine Conservation Society in Phoenix, has been helping with volunteer efforts to contain the spill.

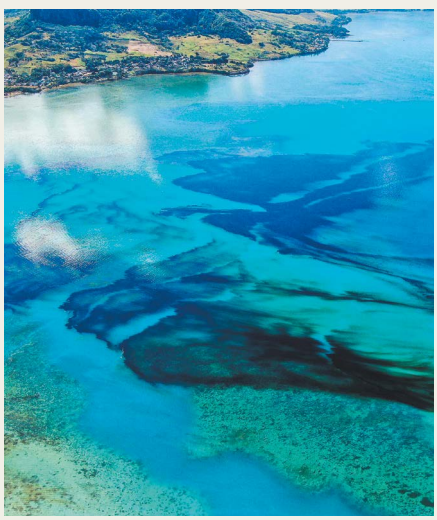
What has been the response to the spill?
Mauritius is not geared up to deal with a catastrophe of this size, so other countries have sent specialists to help. A French team arrived first, from the nearby island Réunion, to erect ocean booms — floating structures that contain the spill. The United Nations sent a team including experts in oil spills and crisis management. Marine ecologists and others have arrived from Japan and the United Kingdom.

Mauritians were also very proactive. In one weekend, we built nearly 80 kilometres of makeshift ocean booms out of cane trash — the leftover leaves and waste from sugar-cane processing.

People worked night and day to stop as much oil as possible reaching the shoreline, where it is more difficult to clean. We managed to contain and remove nearly 75% of the spilt oil. Only a small amount reached the shore. But there's still the issue of water-soluble chemicals that come from the oil, but dissolve into the water and therefore aren't scooped out with the oil that sits on the water's surface.

What ecosystems have been affected?
Unfortunately, there are a lot of environmentally sensitive areas in the region affected. The ship ran aground off Pointe d'Esny and just to the north of Blue Bay Marine Park. These sites are listed under the Ramsar Convention on Wetlands of International Importance as biodiversity hotspots. Ocean currents carried the oil northwards, so fortunately there's none in the Blue Bay Marine Park, but the mangroves on the shoreline north of Pointe d'Esny have been covered. This will definitely have an impact, because mangroves are the nursery of the marine environment.

The Île aux Aigrettes, a small island near the wreck, has also been affected. The island is home to vulnerable pink pigeons (*Nesoenas mayeri*) and other native birds, and to Telfair's skink (*Leiolopisma telfairii*). The oil didn't go onto the island itself, but chemicals might have seeped into the corals.



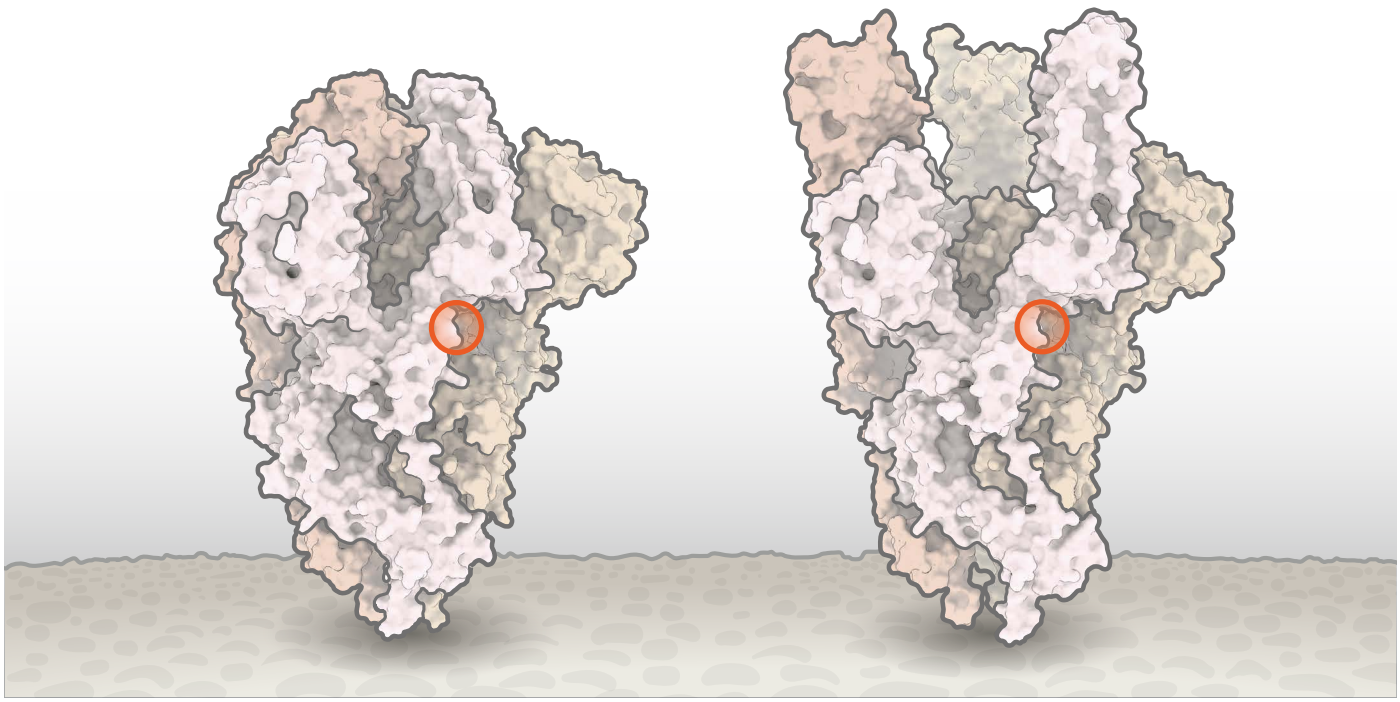
Oil from MV *Wakashio* off Mauritius's coast.

PIERRE DALAIS/EPA-EFE/SHUTTERSTOCK

Are there particular species affected?
It is not one species that could be at risk. It's the whole ecosystem, because of the dispersal of water-soluble chemicals in the water. Filter feeders, such as corals and crustaceans, are probably the first to be affected. We haven't seen lots of animals dying, but we will need to monitor for signs.

Something that is also concerning is that we don't know the possible long-term effects. The oil is a new low-sulfur fuel that is being introduced to reduce air pollution. This is the first time that type of oil has spilled, so there have been no long-term studies on potential impacts.

Interview by Dyani Lewis
This interview has been edited for length and clarity.



SOURCE: STRUCTURAL DATA FROM K. SHEN & J. LUBAN

The spike protein of SARS-CoV-2 has a common mutation (circled) that seems to shift the protein from a closed (left) to an open (right) form.

MAKING SENSE OF CORONAVIRUS MUTATIONS

Different SARS-CoV-2 strains haven't yet had a major impact on the course of the pandemic — but they might in future. **By Ewen Callaway**

When COVID-19 spread around the globe this year, David Montefiori wondered how the deadly virus behind the pandemic might be changing as it passed from person to person. Montefiori is a virologist who has spent much of his career studying how chance mutations in HIV help it to evade the immune system. The same thing might happen with SARS-CoV-2, he thought.

In March, Montefiori, who directs an AIDS-vaccine research laboratory at Duke University in Durham, North Carolina, contacted Bette Korber, an expert in HIV evolution and a long-time collaborator. Korber, a computational biologist at the Los Alamos National Laboratory (LANL) in Sante Fe, New Mexico, had already started scouring thousands of coronavirus genetic sequences for mutations that might have changed the virus's properties as it made its way around the world.

Compared with HIV, SARS-CoV-2 is changing much more slowly as it spreads. But one mutation stood out to Korber. It was in the gene encoding the spike protein, which helps virus particles to penetrate cells. Korber saw the mutation appearing again and again in samples from people with COVID-19. At the 614th amino-acid position of the spike protein, the amino acid aspartate (D, in biochemical shorthand) was regularly being replaced by glycine (G) because of a copying fault that altered a single nucleotide in the virus's 29,903-letter RNA code. Virologists were calling it the D614G mutation.

In April, Korber, Montefiori and others warned in a preprint posted to the bioRxiv server that "D614G is increasing in frequency at an alarming rate"[1]. It had rapidly become the dominant SARS-CoV-2 lineage in Europe and had then taken hold in the United States, Canada and Australia. D614G represented a "more transmissible form of SARS-CoV-2", the paper declared, one that had emerged as a product of natural selection.

These assertions dismayed many scientists. It wasn't clear that the D614G viral lineage was more transmissible, or that its rise indicated anything unusual, they said. But alarm spread fast across the media. Although many news stories included researchers' caveats, some headlines declared that the virus was mutating to become more dangerous. In retrospect,

Montefiori says he and his colleagues regret describing the variant's rise as "alarming". The word was scrubbed from the peer-reviewed version of the paper, published in *Cell* in July[2].

The work sparked a frenzy of interest in D614G. Even those who were sceptical that the mutation had changed the virus's properties agreed that it was intriguing, because of its meteoric rise and ubiquity. For months, that lineage has been found in almost all sequenced samples of SARS-CoV-2 (see 'Global spread'). "This variant now is the pandemic. As a result, its properties matter," wrote Nathan Grubaugh, a viral epidemiologist at the Yale School of Public Health in New Haven, Connecticut, and two colleagues in a *Cell* essay on Korber and Montefiori's findings[3].

So far, the upshot of this work is less clear than Montefiori and Korber's preprint suggested. Some experiments suggest that viruses carrying the variant infect cells more easily. Other work has revealed possible good news: the variant might mean that vaccines can target SARS-CoV-2 more easily. But many scientists say there remains no solid proof that D614G has a significant effect on the spread of the virus, or that a process of natural selection explains its rise. "The jury's out," says Timothy Sheahan, a coronavirologist at the University of North Carolina at Chapel Hill. "This mutation might mean something, or it might not."

Researchers still have more questions than answers about coronavirus mutations, and no one has yet found any change in SARS-CoV-2 that should raise public-health concerns, Sheahan, Grubaugh and others say. But studying mutations in detail could be important for controlling the pandemic. It might also help to pre-empt the most worrying of mutations: those that could help the virus to evade immune systems, vaccines or antibody therapies.

Slow change

Soon after SARS-CoV-2 was detected in China, researchers began analysing viral samples and posting the genetic codes online. Mutations — most of them single-letter alterations between viruses from different people — allowed researchers to track the spread by linking closely related viruses, and to estimate when SARS-CoV-2 started infecting humans.

Viruses that encode their genome in RNA, such as SARS-CoV-2, HIV and influenza, tend to pick up mutations quickly as they are copied inside their hosts, because enzymes that copy RNA are prone to making errors. After the severe acute respiratory syndrome (SARS) virus began circulating in humans, for instance, it developed a kind of mutation called a deletion that might have slowed its spread[4].

But sequencing data suggest that coronaviruses change more slowly than most other RNA viruses, probably because of a 'proofreading' enzyme that corrects potentially fatal copying mistakes. A typical SARS-CoV-2 virus accumulates only two single-letter mutations per month in its genome — a rate of change about half that of influenza and one-quarter that of HIV, says Emma Hodcroft, a molecular epidemiologist at the University of Basel, Switzerland.

Other genome data have emphasized this stability — more than 90,000 isolates have been sequenced and made public (see www.gisaid.org). Two SARS-CoV-2 viruses collected from anywhere in the world differ by an average of just 10 RNA letters out of 29,903, says Lucy Van Dorp, a computational geneticist at University College London, who is tracking the differences for signs that they confer an evolutionary advantage.

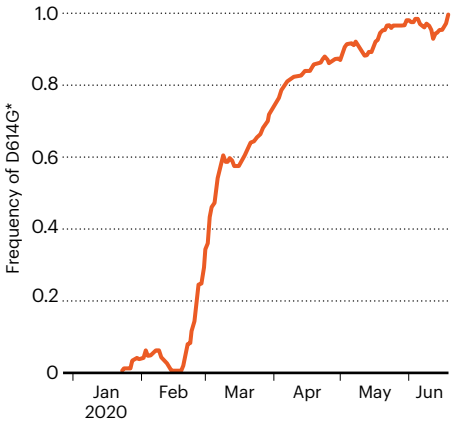
Despite the virus's sluggish mutation rate, researchers have catalogued more than 12,000 mutations in SARS-CoV-2 genomes. But scientists can spot mutations faster than they can make sense of them. Many mutations will have no consequence for the virus's ability to spread or cause disease, because they do not alter the shape of a protein, whereas those mutations that do change proteins are more likely to harm the virus than improve it (see 'A catalogue of coronavirus mutations'). "It's much easier to break something than it is to fix it," says Hodcroft, who is part of Nextstrain (https://nextstrain.org), an effort to analyse SARS-CoV-2 genomes in real time.

"This variant now is the pandemic. As a result, its properties matter."

Many researchers suspect that if a mutation did help the virus to spread faster, it probably happened earlier, when the virus first jumped into humans or acquired the ability to move efficiently from one person to another. At a time when nearly everyone on the planet is susceptible, there is likely to be little evolutionary pressure on the virus to spread better, so even potentially beneficial mutations might not flourish. "As far as the virus is concerned, every single person that it comes to is a good piece of meat," says William Hanage, an epidemiologist at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts. "There's no selection to be doing it any better."

GLOBAL SPREAD

By the end of June, the D614G mutation was found in almost all SARS-CoV-2 samples worldwide.



*Number of samples = 52,292

SOURCE: REF. 8

Faster spread?

When Korber saw the rapid spread of D614G, she thought she might have found an example of meaningful natural selection. The mutation caught her eye because of its position in the spike protein, which is a major target for 'neutralizing' antibodies that bind to the virus and render it non-infectious. And viruses with the mutation were also rising in frequency in more than one part of the world.

D614G was first spotted in viruses collected in China and Germany in late January; most scientists suspect the mutation arose in China. It's now almost always accompanied by three mutations in other parts of the SARS-CoV-2 genome — possible evidence that most D614G viruses share a common ancestor.

D614G's rapid rise in Europe drew Korber's attention. Before March — when much of the continent went into lockdown — both unmutated 'D' viruses and mutated 'G' viruses were present, with D viruses prevalent in most of the western European countries that geneticists sampled at the time. In March, G viruses rose in frequency across the continent, and by April they were dominant, reported Korber, Montefiori and their team[1,2].

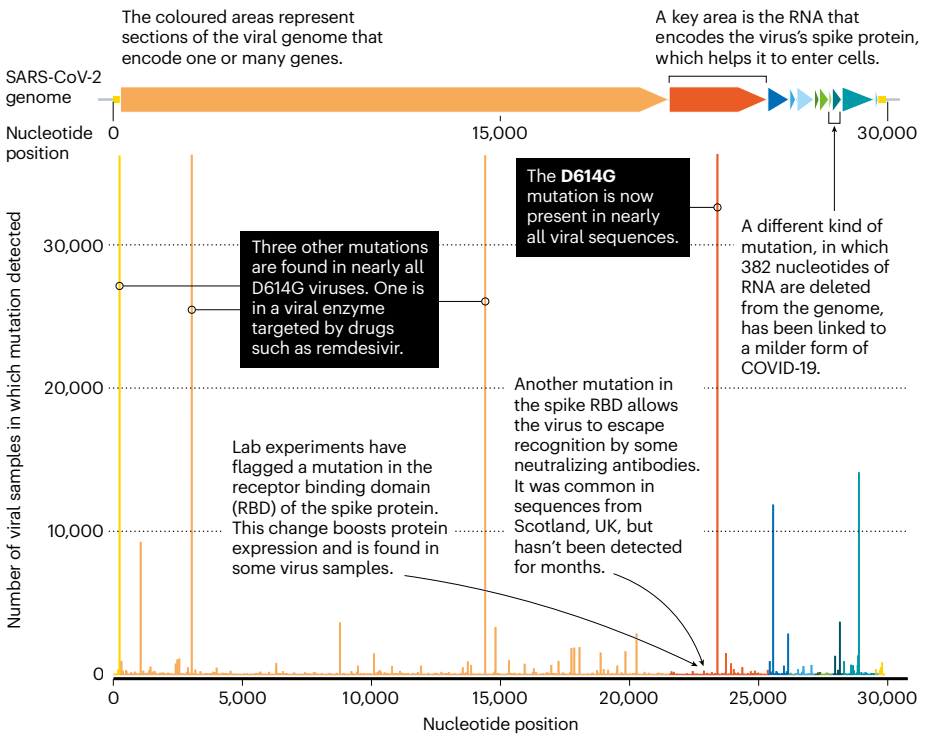
But natural selection in favour of G viruses isn't the only, or even the most likely, explanation for this pattern. The European dominance of G variants could be simply down to chance — if, for instance, the mutation happened to be slightly more common in the viruses that arrived in Europe. A small number of individuals seem to be responsible for most of the virus's spread, and an early, chance tilt in favour of G viruses could explain the lineage's apparent takeover now. Such 'founder effects' are common in viruses, especially when they spread unchecked, as SARS-CoV-2 did in much of Europe until mid- to late March.

Korber and her colleagues tried to rule out a founder effect, by showing in their April preprint[1] that D614G rose to dominance quickly in Canada, Australia and parts of the United States (an exception was Iceland, where G viruses present early in its outbreak were overtaken by D viruses). Analysing hospitalization data from Sheffield, UK, the team found no evidence that viruses carrying the mutation made people any sicker. But those infected with G viruses seemed to have slightly higher levels of viral RNA in their noses and mouths than did those with D viruses.

Many scientists weren't convinced that D614G's rise was remarkable — or all that relevant to the pandemic. "I thought that preprint

A CATALOGUE OF CORONAVIRUS MUTATIONS

Various mutations have been detected in SARS-CoV-2 genomes, including the most prevalent one, D614G. The virus's genetic code has just under 30,000 nucleotides of RNA, or letters, that spell out at least 29 genes. The most common mutations are single-nucleotide changes.



SOURCES: L. VAN DORP *ET AL.* (HTTP://GO.NATURE.COM/3GSRNH6); REFS 2, 11, 12; B. E. YOUNG *ET AL. LANCET* **396**, 603–611 (2020)

was incredibly premature," says Sheahan.

Montefiori says his and Korber's perspective on D614G was shaped by their work on HIV, which has found that even seemingly insignificant mutations can have a profound effect on how the immune system recognizes that virus. "We were alarmed by it, and we need to see if it's having an effect on vaccines," he says.

Rush of lab studies

To examine further whether D614G made the virus more transmissible, Montefiori gauged its effects under laboratory conditions. He couldn't study the natural SARS-CoV-2 virus in his lab, because of the biosafety containment required. So he studied a genetically modified form of HIV that used the SARS-CoV-2 spike protein to infect cells. Such 'pseudovirus' particles are a workhorse of virology labs: they enable the safe study of deadly pathogens such as the Ebola virus, and they make it easy to test the effects of mutations.

The first team to report pseudovirus experiments on D614G, in June, was led by Hyeryun Choe and Michael Farzan, virologists at the Scripps Research Institute in La Jolla, California[5]. Several other teams have posted similar studies on bioRxiv (Montefiori's experiments, and those of another collaborator, appeared in the *Cell* paper[2]). The teams used different pseudovirus systems and tested them on various kinds of cell, but the experiments pointed to the same conclusion: viruses carrying the G mutation infected cells much more ably than did D viruses — up to ten times more efficiently, in some cases.

In laboratory tests, "all of us agree that D to G is making the particles more infectious", says Jeremy Luban, a virologist at the University of Massachusetts Medical School in Worcester. But these studies come with many caveats — and their relevance to human infections is unclear. "What's irritating are people taking their results in very controlled settings, and saying this means something for the pandemic. That, we are so far away from knowing," says Grubaugh. The pseudoviruses carry only the coronavirus spike protein, in most cases, and so the experiments measure only the ability of these particles to enter cells, not aspects of their effects inside cells, let alone on an organism. They also lack the other three mutations that almost all D614G viruses carry. "The bottom line is, they're not the virus," says Luban.

Some labs are now working with infectious SARS-CoV-2 viruses that differ by only the single amino acid. These are tested in laboratory cultures of human lung and airway cells, and in lab animals such as ferrets and hamsters. For labs with the experience and the biosafety capabilities to manipulate viruses, "this is like bread-and-butter kind of work", says Sheahan. The first of those studies, led by researchers at the University of Texas Medical Branch in Galveston, was reported in a 2 September preprint[6]. It found that viruses with the mutation were more infectious than were D viruses in a human lung cell line and in airway tissues, and that mutated viruses were present at greater levels in the upper airways of infected hamsters[6].

Even these experiments might not offer absolute clarity. Some studies show that certain mutations to the spike protein in the Middle East respiratory syndrome (MERS) virus can cause more-severe disease in mice — yet other mutations in the protein show very little effect in people or in camels, the likely reservoir for human MERS infections, says Stanley Perlman, a coronavirologist at the University of Iowa in Iowa City.

The clearest sign that D614G has an effect on the spread of SARS-CoV-2 in humans comes from an ambitious UK effort called the COVID-19 Genomics UK Consortium, which has analysed genomes of around 25,000 viral samples. From these data, researchers have identified more than 1,300 instances in which a virus entered the United Kingdom and spread, including examples of D- and G-type viruses.

A team led by Andrew Rambaut, an evolutionary biologist at the University of Edinburgh, UK, epidemiologist Erik Volz, at Imperial College London, and biologist Thomas Connor at Cardiff University, studied the UK spread of 62 COVID-19 clusters seeded by D viruses and 245 by G viruses[7]. The researchers found no clinical differences in people infected with either virus. However, G viruses tended to transmit slightly faster than lineages that didn't carry the change, and formed larger clusters of infections. Their estimates of the difference in transmission rates hover around 20%, Volz says, but the true value could be a bit higher or lower. "There's not a large effect in absolute terms," says Rambaut.

It's possible that D614G is an adaptation that helps the virus to infect cells or compete with viruses that don't carry the change, while altering little about how SARS-CoV-2 spreads between people or through a population, Rambaut says. "This might be a bona fide adaptation to humans or some human cells," agrees Grubaugh, "but that doesn't mean anything changes. An adaptation doesn't have to make it more transmissible."

Grubaugh thinks that D614G has received too much attention from scientists, in part because of the high-profile papers it has garnered. "Scientists have this crazy fascination with these mutations," he says. But he also sees D614G as a way to learn about a virus that doesn't have much in the way of genetic diversity. "The virologist in me looks at these things and says it would be a lot of fun to study," he says. "It creates this whole rabbit hole of different things you can go into."

He'll have company. Intense study of D614G should help to explain how SARS-CoV-2 fuses with cells, says Luban — a process that might be blocked by drugs or targeted by a vaccine. In an updated version of their pseudovirus experiments posted on bioRxiv on 16 July[8], Luban's team used cryo-electron microscopy to analyse the structure of spike proteins bearing the D614G change. The spike protein is comprised of three identical peptides in an 'open'

or 'closed' orientation. Previous research has suggested that at least two of the three peptides need to be open for the viral particle to fuse with the cell membrane[9], and Luban's team found that viruses carrying the G spike variant were much more likely to be in this state (see 'The mutation that loosens the spike protein'). Computational modelling work by Montefiori and Korber, led by Korber's LANL colleague Sandrasegaram Gnanakaran, came to the same conclusion[10]. "It looks like this molecular machine is primed to go in a way that D is not," Luban says.

No escape from antibodies — yet

Most available evidence suggests that D614G doesn't stop the immune system's neutralizing antibodies from recognizing SARS-CoV-2, as Montefiori had worried. That might be because the mutation is not in the spike protein's receptor-binding domain (RBD), a region that many neutralizing antibodies target: the RBD binds to the cell-receptor protein ACE2, a key step in the virus's entry to cells.

But evidence is emerging that other mutations could help the virus to avoid some antibodies. A team led by virologists Theodora Hatziioannou and Paul Bieniasz, at Rockefeller University in New York City, genetically modified the vesicular stomatitis virus — a livestock pathogen — so that it used the SARS-CoV-2 spike protein to infect cells, and grew it in the presence of neutralizing antibodies. Their goal was to select for mutations that enabled the spike protein to evade antibody recognition. The experiment generated spike-protein mutants that were resistant to antibodies taken from the blood of people who had recovered from COVID-19, as well as to potent 'monoclonal' antibodies that are being developed into therapies. Every one of the spike mutations was found in virus sequences isolated from patients, report Hatziioannou, Bieniasz and their team — although at very low frequencies that suggest positive selection is not yet making the mutations more common[11].

Other scientists are trying to stay ahead of SARS-CoV-2's evolution by predicting which mutations are likely to be important. Jesse Bloom, an evolutionary virologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington, led a team that created nearly 4,000 mutated versions of the spike protein's RBD, and measured how the alterations affected the expression of the spike protein and its ability to bind to ACE2. Most of the mutations had no effect on or hindered these properties, although a handful improved them[12]. Some of these mutations have been identified in people with COVID-19, but Bloom's team found no signs of natural selection for any of the variants. "Probably the virus binds to ACE2 about as well as it needs to right now," he says.

The researchers didn't test whether any of the mutations allow the virus to thwart the action of antibodies, but his team's results suggest that such changes are possible. "It is a possibility, but by no means a certainty, that the virus will acquire mutations that change its susceptibility to antibodies and immunity," says Bloom.

Based on experience with other coronaviruses, that might take years. Studies of common-cold coronaviruses, sampled across multiple seasons, have identified some signs of evolution in response to immunity. But the pace of change is slow, says Volker Thiel, an RNA virologist at the Institute of Virology and Immunology in Bern. "These strains remain constant, more or less."

With most of the world still susceptible to SARS-CoV-2, it's unlikely that immunity is currently a major factor in the virus's evolution. But as population-wide immunity rises, whether through infection or vaccination, a steady trickle of immune-evading mutations could help SARS-CoV-2 to establish itself permanently, says Sheahan, potentially causing mostly mild symptoms when it infects individuals who have some residual immunity from a previous infection or vaccination. "I wouldn't be surprised if this virus is maintained as a more common, cold-causing coronavirus." But it's also possible that our immune responses to coronavirus infections, including to SARS-CoV-2, aren't strong or long-lived enough to generate selection pressure that leads to significantly altered virus strains.

"It is a possibility that the virus will acquire mutations that change its susceptibility to antibodies and immunity."

Worrisome mutations could also become more common if antibody therapies aren't used wisely — if people with COVID-19 receive one antibody, which could be thwarted by a single viral mutation, for example. Cocktails of monoclonal antibodies, each of which can recognize multiple regions of the spike protein, might lessen the odds that such a mutation will be favoured through natural selection, researchers say. Vaccines arouse less concern on this score because, like the body's natural immune response, they tend to elicit a range of antibodies.

It's even possible that the D614G change could make the virus an easier target for vaccines, Montefiori's team found in a study posted to bioRxiv in July[13]. Mice, monkeys and humans that received one of a number of experimental RNA vaccines, including one being developed by drug maker Pfizer in New York City, produced antibodies that proved more potent at blocking G viruses than D viruses.

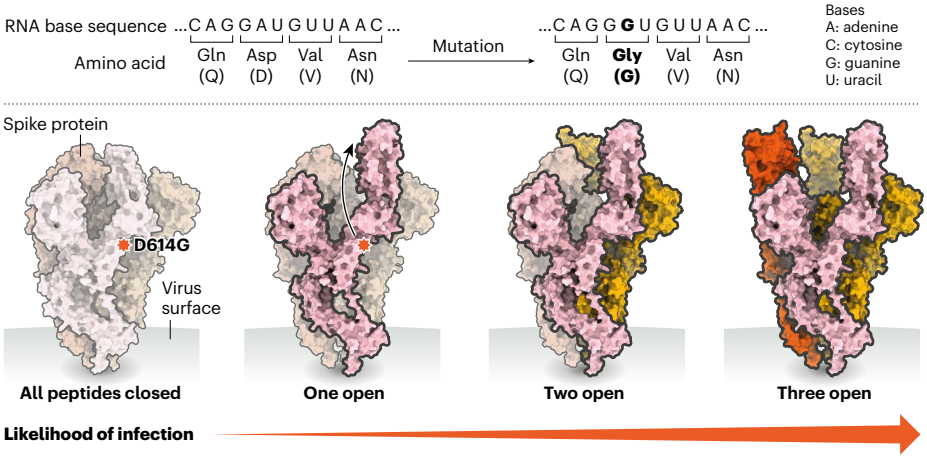
With G viruses now ubiquitous, the finding is "good news", says Montefiori. But as a scientist who has watched HIV mutate to elude many vaccines developed against it, he remains wary of the potential of SARS-CoV-2 to evade humanity's responses. Luban agrees: "We need to keep our eyes open for additional changes."

Ewen Callaway writes for *Nature* from London.

1. Korber, B. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.04.29.069054 (2020).
2. Korber, B. *et al. Cell* **182**, 812–827 (2020).
3. Grubaugh, N., Hanage, W. & Rasmussen, A. *Cell* **182**, 794–795 (2020).
4. Muth, D. *et al. Sci. Rep.* **8**, 15177 (2018).
5. Zhang, L. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.06.12.148726 (2020).
6. Plante, K. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.09.01.278689 (2020).
7. Volz, E. M. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.07.31.20166082 (2020).
8. Yurkovetskiy, L. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.07.04.187757 (2020).
9. Wrapp, D. *et al. Science* **367**, 1260–1263 (2020).
10. Mansbach, R. A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.07.26.219741 (2020).
11. Weisblum, Y. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.07.21.214759 (2020).
12. Starr, T. N. *et al. Cell* **182**, 1295–1310 (2020).
13. Weissmann, D. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.07.22.20159905 (2020).

THE MUTATION THAT LOOSENS THE SPIKE PROTEIN

Spike proteins on SARS-CoV-2 bind to receptors on human cells, helping the virus to enter. A spike protein is made up of three smaller peptides in 'open' or 'closed' orientations; when more are open, it's easier for the protein to bind. The D614G mutation — the result of a single-letter change to the viral RNA code — seems to relax connections between peptides. This makes open conformations more likely and might increase the chance of infection.



SOURCE: STRUCTURAL DATA FROM K. SHEN & J. LUBAN

CRIMEFIGHTING WITH FAMILY TREES

Parabon Nanolabs shot to fame in a controversial field, using DNA and genealogy analysis to catch criminals. Then it had to change tack. **By Carrie Arnold**

It was April 2019 when it all started to fall apart for Parabon Nanolabs. At the time, it was the most famous forensic-genetics company on the planet. From its headquarters in Reston, Virginia, Parabon was helping police to crack cold-crime cases almost weekly, such as the murder of a Canadian couple in 1987 and the case of a young woman who was sexually assaulted and killed in the 1960s.

The company had made its name by comparing suspects' DNA to profiles on genealogy databases and piecing together family trees to track down alleged offenders.

But all those wins had involved long-abandoned cases. Then Parabon helped to solve its first active case, in which a teenage boy had violently assaulted a septuagenarian in a Mormon meeting house in Utah. What could have been the crowning achievement for Parabon ended up stopping the business's meteoric rise overnight.

It was nixed by concerns over privacy. Genealogists at Parabon had been generating leads by sifting through a database of DNA tests called GEDMatch, a free-to-use website that allows users to upload test results in the hope of finding long-lost relatives. At the time, GEDMatch allowed law-enforcement agencies access to the profiles to help solve murders and sexual assaults, unless users specifically opted out. The police, aided by Parabon and companies like it, made new arrests weekly.

But the Utah case was not a murder or a sexual assault — and so was not covered by the website's disclaimer. The assailant had left traces of blood at the scene, and the detective in charge of the case, Mark Taggart, made a personal plea to GEDMatch's founder, Curtis Rogers, for access to the database. When it was granted, Parabon, which had initially refused the case, signed on. The company traced several partial DNA matches to individuals living in the area, and narrowed in on a suspect, a teenaged boy who was a relative of one of them. Taggart made an arrest.

That triggered an immediate backlash from genealogists, privacy experts and the wider public at the violation of GEDMatch's agreement with its users. In response, Rogers required the site's millions of users to specifically opt in to law-enforcement use. Overnight, Parabon lost its lifeblood.

That proved to be a turning point for the company, and for forensic genetic genealogy. In the year since then, the restrictions on GEDMatch's data have forced Parabon to chart a new path forward by returning to one of its earlier business strategies: attempting to use DNA to reconstruct faces. Parabon still offers a forensic genealogy service, but the restrictions have created openings for competitors, which are trying to stake their own claims in the field.

Just as the prominence of forensic genetic profiling has grown, so has its notoriety. Ethicists have raised concerns over China's use of genetic profiling to target the Uyghurs, a predominantly Muslim minority population in the country's northwestern provinces. In the past year, the US government has launched two programmes that have begun taking DNA samples from immigrant detainees and some asylum seekers. The US Department of Justice issued guidelines last November that tried to set boundaries on the use of forensic genetic genealogy, but concerns about police brutality and systemic racism against Black Americans have raised questions as to whether these guidelines provide enough protection to people of colour, who are disproportionately stopped by police and overrepresented in criminal DNA databases. These legal, ethical and social concerns — coupled with Parabon's travails — have left industry experts wondering what's next for forensic genomics.

"Because DNA is so powerful, we tend to see it as a silver bullet," says Yves Moreau, a biologist and engineer at the Catholic University of Leuven in Belgium. But law-enforcement agencies are using databases and techniques

Traces of crime-scene DNA have been matched to suspects using genealogy databases.

JOCHEN TACK/IMAGEBROKER/ALAMY

not designed for solving crimes or generating leads, he says. "It's like a knife — people underestimate just how sharp they can be."

Family ties

In December 2017, genetic genealogist Barbara Rae-Venter got the call that would propel family-tree forensics into the public eye. She was running a business that used GEDMatch to find clients' long-lost relatives when she heard from a California detective who had found some old DNA evidence and was trying to reopen the case of the Golden State Killer, a serial rapist and murderer who committed a string of crimes in the 1970s and 1980s.

Combining DNA samples with family trees is the core of forensic genetic genealogy. The process rests on the simple statistical rules of genetics. A parent and child, or two siblings, share 50% of their DNA. Grandparents and grandchildren share 25%. Even distant relatives share small portions of DNA. This allows consumer genetic-testing companies such as Ancestry in Lehi, Utah, and 23andMe in Sunnyvale, California, to estimate relationships between two individuals who have submitted samples, as far out as fourth cousins (who share a pair of great-great-great grandparents). Anyone can upload the results of their own DNA test to databases such as GEDMatch.

Rae-Venter found two GEDMatch profiles that looked to be second cousins of the suspect, and used that information to work backwards and find their great-grandparents. Then, she moved forward in time to trace their descendants, focusing on California during the time the crimes were committed. After two months, Rae-Venter handed the detective the names of three brothers. DNA from a cigarette discarded by one brother matched the sample, and on 24 April 2018, police arrested Joseph DeAngelo — in the first criminal case to be solved using the technique. (DeAngelo pleaded guilty to multiple counts of rape and murder and was sentenced to life in prison last month.)

> "WE'RE STILL ASKING WHETHER THESE TECHNIQUES ARE SCIENTIFICALLY VALID."

Following DeAngelo's arrest, forensic genetic genealogists such as Rae-Venter and CeCe Moore (who joined Parabon in May 2018) helped to solve similar cold cases at a rapid clip. Although a few ethicists raised concerns about privacy, media coverage of the cases was overwhelmingly positive. "I was actually surprised there wasn't more criticism," says geneticist Ellen McRae Greytak, bioinformatics chief at Parabon.

And then the Utah case hit the media, and the criticism came crashing in.

Active case

Late on Saturday 17 November 2018, 71-year-old Margaret Orlando dialled 911 from a Mormon meeting house in Centerville, Utah. Someone had thrown a rock through a window, climbed in, and attacked her as she was practising the organ, strangling her until she passed out. Taggart was called to the scene, where he found three drops of blood, presumably from her attacker having cut himself on the broken glass. The DNA profile didn't match anyone in state and federal databases, but a chance conversation with a genealogist friend gave Taggart hope: if police couldn't identify the suspect, perhaps they could track down a relative. He reached out to GEDMatch and got permission to use the site.

In the same way that Rae-Venter helped to identify the Golden State Killer, Parabon provided Taggart with three possible names, one of which he recognized right away. The man, who lived near the meeting house, had had several run-ins with the police, and Taggart discovered that he had a 17-year-old nephew living with him — a nephew who matched the description the organist had given.

The next day, Taggart managed to get a DNA sample from a milk carton the suspect had thrown in the rubbish at school. It matched. So did a follow-up swab. Taggart arrested the suspect (whose name was not disclosed as he was a minor) on 24 April 2019 — one year to the day after the arrest of the Golden State Killer. "It was like a puzzle coming together," he says.

With the relief, however, came the publicity. "We were a little surprised at how positive the response was to the Golden State Killer and how negative the response was to this," Greytak says. She points to a study in *PLoS Biology*[1] that found 90% of Americans supported police use of forensic genetic genealogy, and says that a small but vocal group led the outcry against the Utah case.

Ethicist Matthias Wienroth at Northumbria

JOE RAEDLE/GETTY

Migrants detained at the US border. The government takes DNA from some asylum seekers.

University in Newcastle, UK, sees it differently. Wienroth raised privacy concerns about this type of search almost as soon as news of the arrest broke. It's your right to relinquish some of your own privacy by uploading your DNA profile to sites such as GEDMatch, Wienroth says, but these sites also reduce the privacy of some of your distant relatives. Indeed, the proliferation of at-home DNA tests has made some genetic genealogy databases so large that a 2018 *Science* paper[2] estimated that the troves could identify 60% of North Americans of European descent, even if they had never themselves taken one of these tests. Greytak and Armentrout say that they have uploaded their own results to GEDMatch and are untroubled by the idea that they might incriminate a distant relative.

"We're still asking whether these techniques are scientifically valid. No one's talking about failures — all I ever hear about are the successes," Wienroth says. He points to the fact that the California police first chased leads from a different branch of the family tree before they realized their mistake and focused on DeAngelo.

But Greytak doesn't see that as a failure. She says that investigative genetic genealogy was never intended to serve as the final answer in a case. Instead, she sees it as a tool to help law enforcement to generate leads. Those leads — Parabon declined to say precisely how many — evaporated with the changes in GEDMatch's policy, taking one of Parabon's major sources of income with it. To stay afloat, Parabon would have to go back to one of its earliest strategies.

Face value

Steven Armentrout started Parabon in his basement to provide supercomputing services. Parabon's first big breakthrough was in 2011, when the fledgling company applied for a US Department of Defense (DoD) grant to try to reconstruct a person's appearance from their DNA — a technique called DNA phenotyping. The DoD wanted to develop the technology to identify makers of improvised explosive devices from the tiny amounts of DNA left on bombs, but they also knew that law enforcement would be interested. Most labs studying DNA phenotyping look for relationships between changes to individual letters of a person's genetic code, known as single-nucleotide polymorphisms (SNPs), and physical characteristics such as eye or hair colour. But Parabon framed the challenge as a machine-learning exercise. Its plan was to collect a large number of DNA samples and face photographs, and train algorithms to pick out relationships. Parabon got the grant.

> "YOU HAVE A RIGHT TO PRIVACY. YOU ALSO HAVE THE RIGHT NOT TO BE MURDERED OR RAPED."

Its approach worked well with large amounts of high-quality DNA from blood samples and cheek swabs. But forensic samples are often small and degraded. When Armentrout hired Greytak in 2014, the company's first goal was to see whether commercial genotyping arrays could get information from forensic samples. When Parabon sent out its first sample, the lab manager phoned and said it would never work. The chips needed 200 nanograms of DNA.

"In the forensics world, 200 nanograms is a truckload," Armentrout says. Parabon had sent a sample with just 1 nanogram. Everyone involved — including Armentrout and Greytak — was surprised to find that it worked. Parabon says it can now sequence enough SNPs to trace family history and build a face with less than 1 nanogram of DNA. Greytak says that the sequencing runs that use such scant quantities of DNA often leave parts of the genetic code blank because the sample is too degraded or too dilute to read. The company's response was to build proprietary algorithms to anticipate such blank spots in its mathematical models. Greytak says that lower-quality DNA can sometimes mean that predictions are made with less confidence — but that problems are rare.

Parabon's goal was ambitious: rather than just telling police that a suspect had fair hair and green eyes, it wanted to provide a comprehensive analysis of someone's ancestry and a composite facial sketch from a DNA sample. The procedure, dubbed Snapshot, was released in December 2014. Parabon says that since 2018 the police have solved more than 120 cases with the help of their genetic genealogy and phenotyping methods (the company declined to disclose the total number of cases for which they were used, citing ongoing investigations).

Other companies have also developed DNA-phenotyping strategies, including the now-defunct Identitas, which specialized in predicting physical appearance using SNPs, and Illumina, the DNA-sequencing giant in San Diego, California, that spun off its forensics branch into a new company, Verogen, also in San Diego, in 2017.

Several academic labs are also researching DNA phenotyping. At Erasmus University Medical Center in Rotterdam, the Netherlands, Manfred Kayser (once an adviser to Identitas) developed IrisPlex in 2011 to predict eye colour from DNA[3]. Since then, his team has added more SNPs to capture more genetic variation and to add other identifiable characteristics, such as hair colour and texture. The Netherlands police began using Kayser's techniques once they were vetted in the scientific literature. The most famous example was in 2012 when they showed that the rape and murder of 16-year-old Marianne Vaatstra was probably not committed by a member of a refugee settlement located close to where her body was discovered.

Unlike Parabon, Kayser does not attempt to weave together different features to try to recreate a person's face. Instead, he uses the individual traits (say, auburn hair and hazel eyes) as law-enforcement leads. He finds Snapshot to be problematic because the technology hasn't been evaluated in the peer-reviewed literature.

"It's very limited, what we know about the face, and this particular company says they can predict it from DNA. It's pretty bad that they don't publish how they do this and how they validated this," Kayser says. Scientists

have published hundreds of papers about the relationship between specific genetic variants and physical features, Kayser says, but researchers still don't know how these individual traits become a unique human face.

Mark Shriver, a geneticist who researches DNA phenotyping at Pennsylvania State University (Penn State) in University Park, says that because the effects of ancestry on facial appearance are so strong, he suspects that Parabon's data are creating a set of average, generic faces that the company then tweaks to fill in the blanks. Without seeing the data and algorithms the company uses in its machine-learning system, Shriver says, "we don't know whether their ability to estimate a face's appearance is better than chance, or if it's an approximation based on what we know about ancestry".

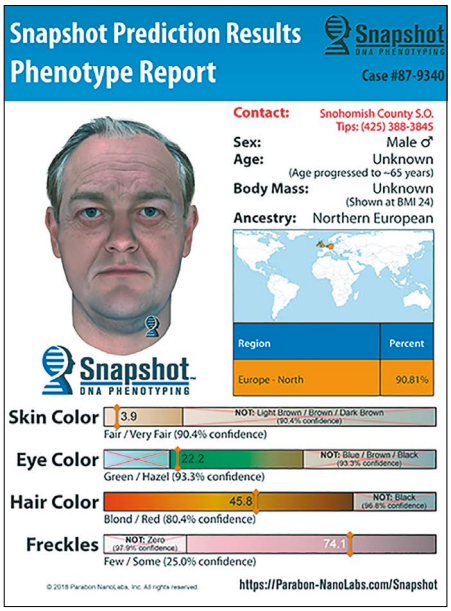
Armentrout says that Parabon doesn't need to know how each gene contributes to appearance in order to create the image of a face; he says the associations between SNPs and faces in the company's database is good enough for its mathematical models, and that police-department satisfaction is all the proof he needs. Just because the firm doesn't publish doesn't mean its method is flawed, Armentrout says. "We're not in business to write papers," he says. "The results speak for themselves." But Shriver says that making an arrest doesn't mean that Snapshot works as Parabon claims. Nor do the police have a rigorous way to show that the Snapshot profile matches their suspect, he says.

SNOHOMISH COUNTY SHERIFF'S OFFICE/NYT/REDUX/EYEVINE

Forensic future

While Parabon was adding DNA phenotyping to its portfolio, other companies, including Verogen and commercial DNA-testing company FamilyTreeDNA in Houston, Texas, began testing the waters with forensic genetic genealogy. Last December, Verogen announced it had bought GEDMatch, which now has 280,000 of its 1.45 million DNA profiles opted in to police searches. Chief executive Brett Williams says that Verogen recognized GEDMatch as the linchpin to forensic genetic genealogy, and wanted to safeguard the company's access. What this means for Parabon and the millions of private GEDMatch users remains to be seen, but Williams says he's committed to striking a balance between privacy and safety. "You have a right to privacy. You also have the right not to be murdered or raped," Williams says. This July, however, GEDMatch was hacked and users' opt-out settings were overridden for a few hours, potentially exposing their data to law-enforcement searches without their consent. In a statement, Verogen said that it had taken down GEDMatch "until such time that we could be absolutely sure that user data was protected against potential attacks".

There have been attempts to gain access to users' profiles through official channels, too. A detective in Orlando, Florida, announced last October that he had obtained a search warrant to use all GEDMatch profiles to try to find relatives from DNA left by a suspect. Genealogy company Ancestry successfully fought against a Pennsylvania search warrant this February. Williams says he will fight against any warrants Verogen receives in the future. In the meantime, the US Department of Justice has issued interim guidelines to help police with their use of forensic genetic genealogy, permitting use of the technology only for serious violent crimes such as rape and murder, and only after other leads have been exhausted. Notably, the document specifies that suspects cannot be arrested on genealogy alone — conventional forensic genetics must be used to provide a conclusive match.



Parabon's Snapshot tool uses DNA to reconstruct faces. This suspect was later convicted of a 1987 murder.

Sociologist Helena Machado at the University of Minho in Braga, Portugal, isn't against law-enforcement use of genetic genealogy or DNA phenotyping, but says she's concerned that work linking genealogy and crime might lead to biases against certain families or ethnic groups. "It might reinforce the idea that there is a higher prevalence of criminality in certain families," she says. An overemphasis on the links between genetics and crime means that researchers could be less likely to focus on the social and economic factors that lead to lawbreaking.

Both Armentrout and Kayser say that DNA technologies could help to reduce police bias by providing concrete evidence to bolster eyewitness accounts, and that DNA phenotyping could decrease racial profiling by providing more details on a potential suspect's appearance to police.

But sociologist Amade M'charek at the University of Amsterdam says this thinking is naive, especially given the incidence of police brutality against people from racial minorities. "If we don't know the individual, often all we see is race," she says.

M'charek's concerns are not unfounded: these technologies are already being used to target and discriminate against people from minority groups, Moreau says. The US Department of Homeland Security announced in January that its Immigration and Customs Enforcement (ICE) division had launched a pilot programme to collect DNA from immigrant detainees and upload the resulting sequences to the Federal Bureau of Investigation's's official forensic DNA database, the Combined DNA Index System (CODIS). The initiative joined last year's announcement that homeland security would be using 'rapid DNA technology' to test whether families applying for asylum were relatives. (ICE did not respond to requests for comment.)

In China's northwest, officials are using genetic ancestry to identify members of the Uyghur minority group. In July 2017, as part of China's Physicals for All programme, the government began collecting iris scans, fingerprints and DNA of everyone between the ages of 12 and 65 in the Xinjiang Uyghur Autonomous Region. The programme has been criticized by human-rights groups. Dispatches from Xinjiang from the non-governmental organization Human Rights Watch in New York City, reported that more than one million Uyghurs have so far been placed in detention camps. "When you give any authority such important information and such strong leverage against individuals, you start to worry very, very much about the shape society's going to take," Moreau says. "You put people in a database because you want to control them." Some Chinese scientists, says Moreau, are also working to turn Uyghur DNA into facial portraits, just as Snapshot does. Parabon says it is not involved in the Chinese research.

Despite the controversy over the Utah case — or perhaps because of it — Rogers is bullish about the future of genetic techniques in forensics. "I think that in time — and probably not very long — people will accept that law-enforcement use of genetic genealogy is there and not to be feared," he says.

For his part, Taggart doesn't regret using GEDMatch. The suspect he narrowed in on pleaded guilty and is still in detention, and Taggart is confident that his community is safer that way. "I believe that Curtis Rogers doing this for us saved a life."

Carrie Arnold is a science journalist based near Richmond, Virginia.

1. Guerrini, C. J., Robinson, J. O., Petersen, D. & McGuire, A. L. *PLoS Biol.* **16**, e2006906 (2018).
2. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. *Science* **362**, 690–694 (2018).
3. Walsh, S. *et al. Forensic Sci. Int. Genet.* **7**, 98–115 (2011).

Books & arts

MARK RALSTON/AFP/GETTY

Health-care workers in a 'White Coats for Black Lives' protest in California in June.

Weapons for when bigotry claims science as its ally

As COVID-19 reveals the toll of discrimination, racism and inequality, a book skewers genetic reductionism.

By Alondra Nelson

This is a year of reckonings. Chief among them: communities have been forced to face the injustices laid bare by the yawning racial and ethnic disparities in illness and death caused by COVID-19 the world over.

Predictably, even the data that shine some light on these inequalities remain wanting. In the United States, the Centers for Disease Control and Prevention withheld national-level data about the disproportionate impacts of COVID-19 on Black, Latinx and other people until threatened with a lawsuit. In the United Kingdom, a government agency removed nearly 70 pages of community-based research from a report that pointed to structural causes of unequal disease toll on Black, Asian and minority ethnic groups.

Still, there is much we do know. The extra burden borne by under-resourced and marginalized communities globally is plain. In the United States, for example, Black residents in the state of Maine reportedly comprise nearly 21% of those infected with COVID-19, despite being just 1.4% of the population. People of Pacific Islander descent, including Native Hawaiians, in Los Angeles County, California, have an infection rate six times that of their white neighbours. Black, Bangladeshi and Pakistani communities in the United Kingdom experience rates of infection with the new coronavirus up to twice those of white communities, and are more likely to become severely ill with the disease.

These data show trends in societally constructed categories; they do not explain how the trends arise. How, then, can we account for tragic losses in groups as distinct as Roma communities in Greece, Indigenous Yanomami in Brazil, and Somali immigrants in Norway? The common inheritance of these diverse populations is the lived experience of discrimination, racism and inequality. Yet, even now, some people prefer to suggest that these health disparities are driven by genetics. It is a wearily, tragically familiar line of reasoning.

An indictment of this sort of genetic reductionism is Adam Rutherford's book *How to Argue with A Racist*. Although it does not deal with genetics in medicine or public health, its efforts are urgently relevant to the present moment. A science broadcaster (and former head of multimedia at *Nature*) trained in genetics, Rutherford parses claims about the purported relationship between DNA and race. His stated aim? To use the "weapon" of scientific fact to vanquish the myth that racism is "grounded in science".

Rutherford's battle plays out in four acts, spanning appearance, ancestry, athleticism and intelligence. Dismantling racist falsehoods that masquerade as truths, he returns to several themes: DNA data are over-interpreted; the environment is under-appreciated; and human genetic difference is "wickedly complicated", is often unpredictable and bears little allegiance to socially constructed, politically significant demographic categories.

Countering the myth that human physical appearance has any predictable relationship with genetics, Rutherford shows that differences in skin colour occur across a wide geography that has little relationship to common ideas of race. He argues, in effect, that from a genetic perspective, skin colour is only skin deep. It is, he writes "a very bad proxy for the total amount of similarity or difference between individuals and between populations". Referencing the work of geneticist Sarah Tishkoff, he reminds us that there is more genetic diversity on the African continent than in the rest of the world, and that this diversity extends to pigmentation. "DNA is a

bewilderingly inscrutable predictor of skin colour," he concludes.

The myth that DNA and genetic genealogy are reliable registers of ancestry, kinship or 'racial purity', Rutherford skewers as nonsense. For example, he notes that human migration patterns do not abide by sociopolitical constructions of country or nation. He joins social scientists (me included) who have been pointing out for years that direct-to-consumer genetics are as much about contemporary genealogical aspirations as about the past.

He engages the research of sociologists Aaron Panofsky and Joan Donovan, who have studied avowed white nationalists and white supremacists keen to demonstrate their notionally 'pure' European ancestral origins. Depending on whether the test results they receive confirm or contradict their hopes, they adopt, reinterpret or reject the data. The "same warping of science", Rutherford writes, "fuels both racists and typical hobbyist genealogists".

Rutherford does survey the history of eugenics. But he does not acknowledge that having this contemptible field at its foundations might prevent genetics ever being the anti-racist ally he hopes. This tension looms over his battle with two other myths: the supposed athletic superiority of people of African descent, and the supposed intellectual prowess of those of Jewish descent.

Sports send genetic determinism into overdrive. One or a very few identified genetic variants are 'fetishized' and made proxies for individuals and entire communities. And racial theories of athleticism are ridiculously inconsistent, offered as explanations for a dizzying array of skills, from swimming to sprinting. For example, it is not some genetic lack of buoyancy — as folk logic would have it — that makes it less likely for African Americans to be competitive swimmers, Rutherford reminds us. Rather, swimming pools were part of the 'Jim Crow' system of US racial apartheid that lasted well into the 1960s. (Anti-Black violence continues to keep these codes in effect, as evidenced by a 2015 viral video showing a Black teenage girl being viciously attacked by police in McKinney, Texas, for attempting to use a community swimming pool with her friends.)

As for the myth of racial correlation with IQ, Rutherford stresses the importance of environmental factors in driving variable measures across populations. This is the weakest part of the book, because Rutherford thoroughly dismantles the case for IQ, yet clings to a belief that it has a definitive basis in evidence. Noting the many confounding factors that contribute to cognitive performance, he writes, "IQ is a single number, but intelligence is not a single thing". This uncertainty registers throughout the chapter. Although "heritability" — which can include influences ranging from shared upbringing to intertwined social networks to biology — might contribute to measures of IQ, the role of genetics remains elusive. And I would add that it has been demonstrated repeatedly that this "single number" encapsulates class assumptions, educational access and other inequities and, that it is malleable on the basis of the presence or absence of these.

ADRIANO MACHADO/REUTERS

Children from the Yanomami Indigenous group don protective masks in Brazil.

Existential moment

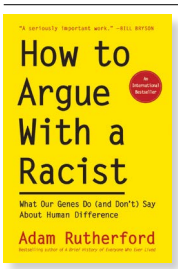
The aim of Rutherford's book is noble, and he mostly succeeds in his endeavour. He deploys genomic variation and unpredictability against those who make claims of static characteristics or who assert bunk about 'racial purity'. He draws on the complexity of gene–environment interactions to bludgeon narrow, incorrect determinism. He highlights the social and political shaping of genetic claim-making.

"A writer never knows what kind of world a book will land in, what will change around the words on the page," wrote novelist and journalist Hari Kunzru in *The New York Review of Books* in July. Rutherford wrote his book before the pandemic, in the context of rising nationalist politics inherently tied to the most regressive ideologies of ancestry, feeling it his duty to contest racism with facts, "especially if bigotry claims science as its ally".

It is published in an existential moment of suffering and death, of global outrage over the killing of unarmed Black people by police, of a reckoning with the legacy of racial slavery and colonialism. Many nations have seen an efflorescence of anti-racist reading lists. Rutherford's book is rightfully on them.

But like many such volumes, after reading, the question remains: in this moment, could arguing the facts, even with Rutherford's compelling narrative and nuance, possibly be enough? Rutherford himself admits: "Arguing with racists with conspiracy mindsets about science is a fairly fruitless endeavour, and exhausting".

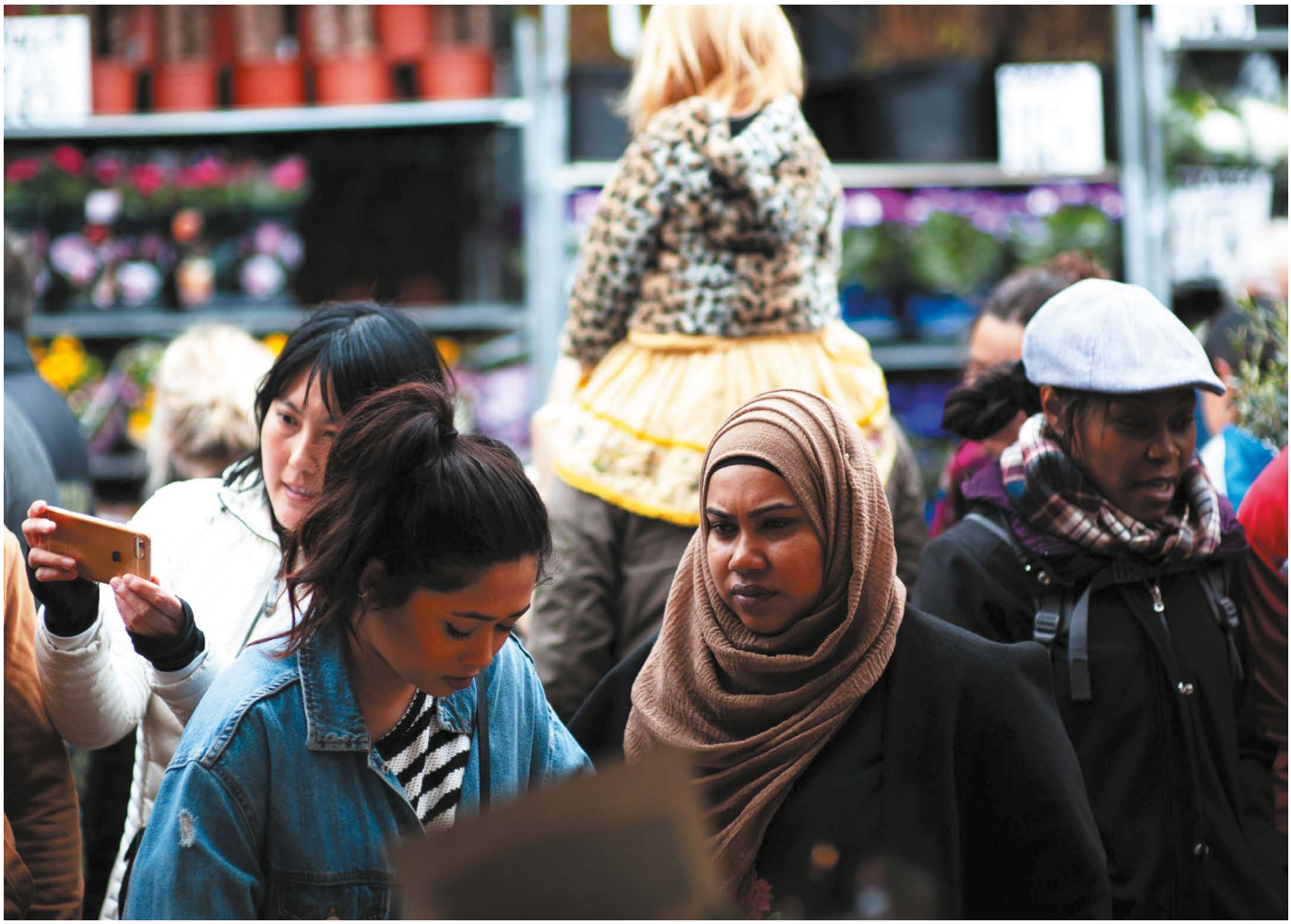
This is a moment for deeds, not words. To topple the edifice of structural racism that produces 'excess' death in the context of COVID-19 and of life generally will take urgent social, political and economic action, from court rooms to clinics, lecture halls to voting booths.



How to Argue With a Racist: What Our Genes Do (or Don't) Say About Human Difference
Adam Rutherford
The Experiment (2020)

Alondra Nelson is the Harold F. Linder Chair in the School of Social Science at the Institute for Advanced Study in Princeton, New Jersey. Her books include *The Social Life of DNA* and *Genetics and the Unsettled Past*.

Comment



ALAMY

Customers at Columbia Road Flower Market, UK.

Don't ignore genetic data from minority populations

Chief Ben-Eghan, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V. Grant, Charles Dupras & Simon Gravel

Efforts to build representative studies are defeated when scientists discard data from certain groups. Instead, researchers should work to balance statistical needs with fairness.

Geneticists have known for more than a decade that their focus on people with European ancestry exacerbates health disparities[1]. A 2018 analysis of studies looking for genetic variants associated with disease found that under-representation persists: 78% of study participants were of European ancestry, compared to 10% of Asian ancestry and 2% of African ancestry. Other ancestries each represented less than 1% of the total[2]. Several projects, such as H3Africa[3], are starting to increase participation of under-represented groups, both among participants and among researchers. Large biobanks assembled in Europe and North America, combining biological samples with health-related data, also set sampling targets to increase diversity[4,5,6].

But even when data from minority groups are available, many researchers discard them[7].

Although there can be valid reasons to restrict analyses to a particular population, discarding such data by default is ethically problematic: it worsens under-representation and negates participants' efforts to contribute to research.

Funding agencies have taken steps to improve the diversity of participants who are recruited for studies — notably, this has led to better representation of women in clinical trials since the 1990s. But agencies have less control over researchers' decisions of what to analyse. Scientists are pulled towards statistical convenience and publishing incentives, which can both conflict with the collective goal of greater equity.

Here we suggest that an approach used in health care can help researchers to make analysis decisions that are ethically as well as scientifically sound.

Ruled out

To estimate how often minority data are excluded, we examined publications that used data from either the UK Biobank (UKB; which contains material from 502,655 individuals) or the US Health and Retirement Study (HRS; 12,454 individuals). Both biobanks support genome-wide association studies (GWAS). These scan data from thousands of participants to find genetic variants associated with disease.

To compare the criteria researchers used to include or exclude data types across studies, we distinguished between participants from majority (MAJ) and minority (MIN) groups in the United States and the United Kingdom. We used MAJ regardless of whether a study focused on self-declared ethnicity, such as 'white', or on the location of an individual's ancestors, such as 'European ancestry'. We used MIN to refer to all other individuals, including those of mixed ancestry or ethnicity. This coarse labelling helps to describe how data were used in statistical analyses, and does not imply that either group is uniform. We counted MIN data as 'included' if any analysis reported linking traits or diseases to genotypes in the relevant samples.

First, we reviewed 21 articles from the GWAS catalogue (www.ebi.ac.uk/gwas) that contained the keywords 'UK biobank' (see Supplementary information). Twenty restricted their analysis to only MAJ individuals in the UKB database (two of these also analysed data from a broader range of ancestries in other databases). We also queried online repositories and randomly sampled another 20 GWAS that used UKB data. Only one used MIN data. Finally, we reviewed 17 GWAS listed on the HRS online publications list. Here, only six studies limited analysis to MAJ populations, perhaps because the proportion of MIN participants in the US biobank (24%) was higher than in the UK one (5%).

Overall, 45 of 58 studies in our sample excluded MIN data. If we weight representation by the number of times data from an individual were actually analysed, MIN representation in the UKB falls to 0.06% (see 'Left out'; details are in Supplementary information). This problematic situation will surprise few genetics researchers[6,7].

> **"By omitting data, scientists squander an opportunity to build useful knowledge about minority populations."**

Both the UKB and the HRS made efforts to represent their national populations. However, including individuals from minority groups in data cohorts but not in analyses can be seen as de facto tokenism. Unused data do not help under-represented groups.

Why exclude?

Of the 45 studies that excluded data, 31 gave no reason. The remaining 14 studies provided 15 explanations for exclusion.

The most common explanation was fear of confounding (11/15). If a genetic variant happens to be more common in an ancestry group, and that group happens to have a higher rate of a particular trait, there will be a correlation between having the variant and having the trait. An example is childhood asthma, which is influenced by both genetic and environmental factors. Researchers might confuse the correlation as evidence that this variant causes childhood asthma. Although statistical methods to avoid confounding exist, they are not foolproof, and confounding is a legitimate concern[7].

It is not necessary to exclude data to reduce the risk of confounding. Data from different groups can simply be analysed separately. However, because samples from minority populations are so much smaller, they have less statistical power and are therefore less likely to reveal new genetic associations.

This lack of power was the second-most-cited reason for exclusion (3/15). An underpowered study can be seen as a waste of time because it might not yield statistically significant results. Because finding a genetic association can be enough to garner a publication, adding analysis of other populations comes at a cost. It takes time, makes the manuscript more complicated, gives reviewers one more thing to criticize, and so could delay publication.

Grounds for inclusion

There is value in data from minority populations.

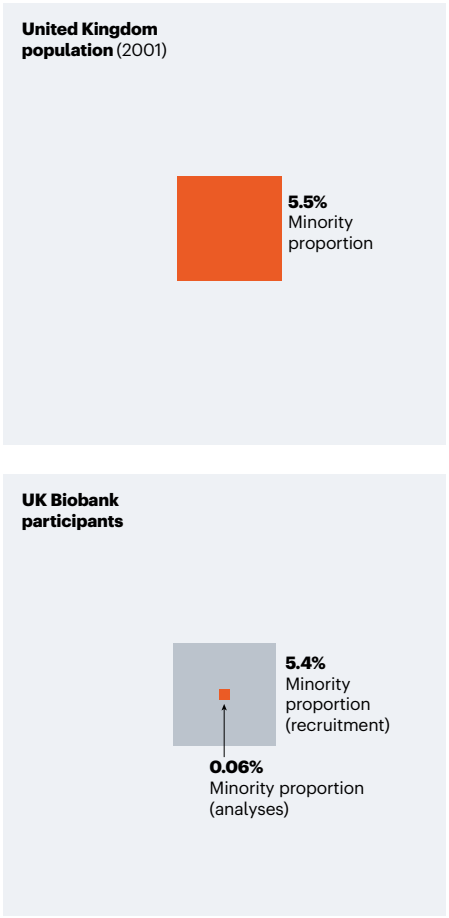
As part of a study on asthma, we performed a genome-wide association study for eosinophil cell counts. (Eosinophils are a subset of white blood cells and are often elevated in individuals with asthma.) We did three separate analyses. One was of the majority (MAJ) population; two were of the minority (MIN) populations defined using the UK Biobank self-reported ethnicity categories (participants who identified as Black or Black British, and those who identified as Asian, Asian British or Chinese).

The MAJ analyses identified 432 genetic loci (1,510 independent genetic variants). The two MIN analyses independently identified 3 loci (at genome-wide significance, $P \leq 5 \times 10^{-8}$), all of which were identified in the MAJ analysis. The MIN analysis enabled validation of more than one-quarter of the identified variants in the MAJ population at nominal significance ($P = 0.05$). It also showed overall consistent results across ethnicities, except for one variant that showed nominal significance, but opposite effects in Asian, Asian British and Chinese populations, relative to the MAJ analysis. Without further evidence, this variant should probably not be used to predict genetic risk outside Europe. (See Supplementary information for details.)

These analyses took 10 hours of computing time as well as some forethought. This is insignificant compared with the cost of accessing the data. Evidence of association for the millions of variants we tested can now be compared across populations and can be made available for meta-analyses. Such data are particularly important for studying minority populations, when samples in individual cohorts might lack statistical power.

LEFT OUT

UK biobank recruitment reflected diversity (in 2001; ref. 11). Analyses do not.



SOURCE: ANALYSIS BY C. BEN-EGHAN *ET AL.*

Just one study explicitly mentioned following methods from past publications as grounds for exclusion (1/15), but we suspect that this is common. There are good reasons to follow precedent: using standard analytical pipelines reduces development cost and the need for extensive validation and explanation.

Together, these three reasons drive researchers to discard data from MIN populations.

Lost opportunity

By omitting data, scientists squander an opportunity to build useful knowledge about minority populations. If researchers perform GWAS on populations of European ancestry, they can often use previously published results in the form of summary statistics to strengthen their findings. Because summary statistics present little privacy risk to participants, they can usually be downloaded freely in just a few minutes. Doing the same comparison with MIN population data that have not been previously reported requires accessing individual-level information. This involves obtaining institutional ethics approval, requesting data access from the cohort, plus cleaning and processing data — all before finally performing GWAS. This can take months. If MIN data are not analysed alongside MAJ data, they might never be used.

When done as part of the primary study, by contrast, MIN analyses add little cost and can be informative (see 'Grounds for inclusion').

Four criteria

Analysing MIN data is important for equity and discovery. But how should we weigh that against the immediate, individual burden of statistical analysis and delayed publication? General rules that apply to all studies are hard to define, but there is an approach that should help.

Over the past two decades, governments and ethicists have leant on a framework called accountability for reasonableness (A4R) to help allocate scarce resources in health care, such as new or expensive treatments. A4R recognizes that individuals in a pluralistic, democratic society give different weight to different considerations, and so might never agree on broad principles. Instead, A4R focuses on the decision-making process itself, and sets out criteria that encourage fairness and legitimacy[8]. In short, reasons for decisions should be transparent and relevant. Adherence to these criteria should be enforced and measured in a way that adapts to new information.

The A4R criteria suggest small changes in analysis and publication conventions that would improve fairness and accountability.

Transparency. In their publications, researchers should state reasons for excluding participant data. More generally, they should explain design and analysis choices that have the potential to worsen inequalities.

Relevance. The stated reasons for exclusion should explain how the decision sought to best serve society, given the real-world constraints of research. Reasons such as fear of confounding, limited power and precedent might not meet this requirement if they can be circumvented by a particular analysis method (using stratified or meta-analysis, for instance). Barring more compelling reasons, we recommend that researchers compute association statistics for MIN populations and report them as part of the primary study.

Enforcement. We propose that journals mandate that submitted manuscripts justify any exclusion of participant data in analyses. Forms should ask reviewers whether relevant reasons were provided.

The goal is not to turn reviewers into moral arbiters. Rather, they should simply assess whether the reasons provided are relevant to the analyses under review. This modest requirement would encourage analyses to be more inclusive, foster broader discussion about legitimate grounds for exclusion and clarify expectations for authors.

Importantly, reviewers should not require results of analyses of MIN and MAJ populations to be consistent. Discrepancies should be discussed, but forcing researchers to explain all observations would prevent useful results from being shared.

Revisions. How researchers assess transparency and relevance should change with society and methodology. Our recommendations that data from MIN populations be analysed by default might become moot if sufficient data become available in cohorts that focus on under-represented groups[9]. The field might also move to a model in which specialized teams analyse MIN data across multiple phenotypes (see, for example, https://pan.ukbb.broadinstitute.org). This would change both the costs and benefits of performing subsequent analyses of MIN data. It could reduce the impetus for analysis by individual studies while providing tools that reduce the analysis burden and risk of confounding for subsequent researchers.

Statistical analyses that are more inclusive cannot overcome fundamental inequities in representation among study participants, let alone solve the broader issues of equity and data sovereignty[10]. But they are a step in the right direction. By acknowledging the tension between ethical and practical considerations, researchers in genetics and other fields can hold themselves accountable for making scientific advances more efficient and more fair.

The authors

Chief Ben-Eghan, **Rosie Sun**, **Hans Markus Munter**, **Alex Diaz-Papkovich**, **Audrey V. Grant** and **Simon Gravel** are genomics researchers and **Charles Dupras** is an ethicist at McGill University in Montreal, Quebec, Canada. **Jose Sergio Hleap** is a high-performance-computing technical consultant in bioinformatics at SHARCNET in Guelph, Ontario, Canada.
e-mail: simon.gravel@mcgill.ca

1. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. *Nature* **475**, 163–165 (2011).
2. Sirugo, G., Williams, S. M. & Tishkoff, S. A. *Cell* **177**, 26–31 (2019).
3. Mulder, N. *et al. Pharmacogenomics Pers. Med.* **11**, 59–66 (2018).
4. Ollier, W., Sprosen, T. & Peakman, T. *Pharmacogenomics* **6**, 639–646 (2005).
5. Sonnega, A. *et al. Int. J. Epidemiol.* **43**, 576–585 (2014).
6. Tutton, R. *Race/Ethnicity: Multidiscip. Glob. Contexts* **3**, 75–95 (2009).
7. Peterson, R. E. *et al. Cell* **179**, 589–603 (2019).
8. Daniels, N. & Sabin, J. E. *Br. Med. J.* **337**, a1850 (2008).
9. Wojcik, G. L. *et al. Nature* **570**, 514–518 (2019).
10. Fox, K. *N. Engl. J. Med.* **383**, 411–413 (2020).
11. Fry, A. *et al. Am. J. Epidemiol.* **186**, 1026–1034 (2017).

Supplementary information accompanies this article (see go.nature.com/2dwmgka).

News & views

Ageing

Molecules in old blood promote cancer spread

Hai Wang & Xiang H.-F. Zhang

A molecule produced by the metabolism of proteins and fats has been found to accumulate in the blood of older people, and to endow cancer cells with the ability to spread from one site in the body to others. **See p.283**

As we get older, the risk that we will develop cancer increases, because we accumulate genetic mutations and are continually exposed to cancer-causing substances[1]. Most cancer-causing agents are found in the environment, but some are produced by our own bodies. Gomes *et al.*[2] report on page 283 that methylmalonic acid (MMA) — a by-product of protein and fat digestion — can accumulate in the blood with age, and might promote the spread of tumours.

Methylmalonic acid is produced in cells in very small amounts[3]. Usually, it becomes linked to the molecule coenzyme A to form methylmalonyl-CoA, and is converted to succinyl-CoA in a reaction that involves vitamin B_{12} as a cofactor. Succinyl-CoA subsequently enters the TCA cycle — a series of chemical reactions that are a key part of energy production in the cell.

In some diseases, the body fails to metabolize MMA efficiently, leading to its toxic accumulation in the blood. For instance, the metabolic disorder methylmalonic acidaemia is characterized by the failed conversion of methylmalonyl-CoA to succinyl-CoA, owing to genetic defects in key enzymes (such as methylmalonyl-CoA mutase) or to vitamin B_{12} deficiency[1].

Gomes *et al.* report that MMA levels are significantly higher in the blood of healthy people over the age of 60 than in those under 30. The elevated level of MMA had not caused ill health in the individuals studied. However, the authors found that treating human cancer cells with serum from the blood of the older group, or with high concentrations of MMA, led them to adopt characteristics of metastatic cancer cells — those that can spread from a primary tumour to seed cancers elsewhere in the body. These characteristics include a loss of cell–cell attachment and an increase in mobility. When injected into mice, the cells formed metastatic tumours in the lungs.

The researchers demonstrated that the presence of large lipid structures in 'old' blood serum was also key to its ability to induce metastatic characteristics in cells. Removing these structures from blood prevented MMA from entering cells, indicating that MMA is in complex with a large lipid. The identity of this lipid structure, and the mechanism by which it helps MMA to enter cells, remains to be determined.

> **"The authors' results should stimulate more interest in the relationship between protein intake and age-associated cancer risks."**

Gomes and colleagues next asked what molecular changes MMA triggers in cells. The authors examined the gene-expression profiles of cells treated with MMA, and compared them with those of untreated cells. One of the genes most highly upregulated in response to MMA was *SOX4*, which encodes a transcription factor involved in the regulation of embryonic development and cancer progression[4]. The authors demonstrated that repressing *SOX4* expression blocked the cancer-cell response to MMA, and prevented the formation of metastatic tumours in mice that received injections of cancer cells treated with old serum. Thus, MMA indirectly induces an increase in the expression of *SOX4*, which in turn elicits broad reprogramming of gene expression and subsequent transformation of cells into a metastatic state (Fig. 1).

Gomes and colleagues' work implies that lipids have dual roles in MMA-driven metastases: first, in the form of the fatty acids from which MMA derives; and second, as large lipids that help MMA to cross cell membranes. Levels of the lipid cholesterol increase between puberty and the age of 50 or 60 (ref. 5) — overlapping with the rise in MMA levels in the blood. It is possible that the lipidic structures observed in the current study involve cholesterol. If so, anti-cholesterol treatments might reduce levels of MMA and slow its entry into cells.

Why does MMA increase with age? Levels of vitamin B_{12} decrease with age, and deficiency in that vitamin is linked to an accumulation of MMA. However, the authors found no reverse correlation between levels of these two molecules in their study participants. Therefore, B_{12} deficiency is unlikely to be the main reason for MMA accumulation. Another potential culprit is protein. A low-protein diet can reduce the substrates for MMA formation[6], and might

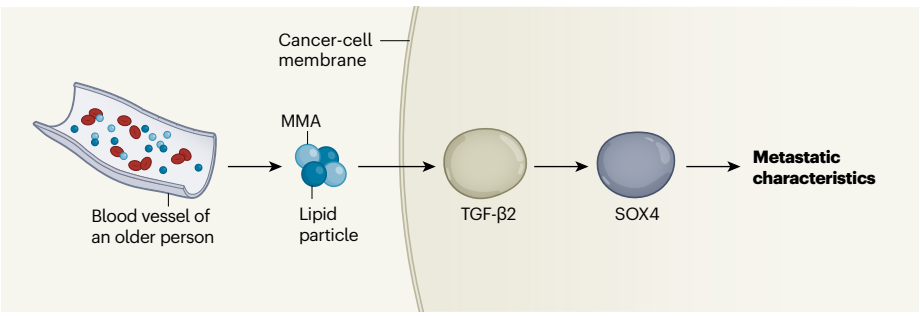


Figure 1 | Methylmalonic acid (MMA) and cancer. MMA is produced during digestion of proteins and fats. Gomes *et al.*[2] report that levels of MMA are elevated in the blood of people over the age of 60, compared with those under 30. The group provides evidence that large lipid structures help MMA to enter cancer cells from older blood vessels. Through unknown pathways, MMA promotes expression of the proteins TGF-β2 and SOX4. In turn, SOX4 drives global gene-expression changes that enable cells to take on the characteristics of metastatic cells, which spread cancer around the body.

enhance anticancer immune responses[7]. In addition, high protein intake significantly increases the risk of death from cancer in people aged 50 to 65 (although the opposite correlation is seen in people over 65)[8]. Given these previous observations, Gomes and colleagues' work should stimulate more interest in the relationship between protein intake and age-associated cancer risks.

All the people in this study who had high plasma levels of MMA seemed to be cancer-free, suggesting that the effects of MMA are specific to cancer spread in the body, rather than to initial cancer formation. Cancer initiation and spread are distinct processes that involve different molecular mechanisms[9]. If future studies can confirm that MMA specifically affects metastasis in humans in the same way that Gomes *et al.* have demonstrated it does *in vitro* and in mice, this molecule will stand apart from many previously known ageing-related causes of cancer, including environmental factors and genetic mutations. Further investigation into the timing of MMA's effects could then inform the optimal timing for therapeutic use of MMA-blocking agents, if they become available.

A final question is how MMA stimulates gene-expression changes associated with metastasis at a molecular level. The authors hypothesized that MMA activates transcription of the gene *TGF-β2*; this gene is part of a TGF-β signalling pathway that, in turn, promotes *SOX4* expression. But how MMA enhances the transcription of *TGF-β2* remains to be seen.

Answers to these questions will further our understanding of metabolic changes and their roles in cancer development. Regardless of the answers, Gomes and colleagues' study has broadened our view of cancer risk factors, by drawing attention to the role of metabolism in ageing-associated cancer progression.

Hai Wang and **Xiang H.-F. Zhang** are in the Lester and Sue Smith Breast Center, the Dan L. Duncan Cancer Center and the Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. **X.H.-F.Z** is also at the McNair Medical Institute, Baylor College of Medicine.
e-mail: xiangz@bcm.edu

1. Anisimov, V. N. *Crit. Rev. Oncol. Hematol.* **45**, 277–304 (2003).
2. Gomes, A. P. *et al. Nature* **585**, 283–287 (2020).
3. Chandler, R. J. & Venditti, C. P. *Mol. Genet. Metab.* **86**, 34–43 (2005).
4. Lourenço, A. R. & Coffer, P. J. *Trends Cancer* **3**, 571–582 (2017).
5. Félix-Redondo, F. J., Grau, M. & Fernández-Bergés, D. *Aging Dis.* **4**, 154–169 (2013).
6. Baumgartner, M. R. *et al. Orphanet J. Rare Dis.* **9**, 130 (2014).
7. Rubio-Patiño, C. *et al. Cell Metab.* **27**, 828–842.e7 (2018).
8. Levine, M. E. *et al. Cell Metab.* **19**, 407–417 (2014).
9. Hanahan, D. & Weinberg, R. A. *Cell* **144**, 646–674 (2011).

This article was published online on 19 August 2020.

Electrical engineering

A cool design for hot microchips

Tiwei Wei

Miniaturized electronic devices generate a lot of heat, which must be dissipated to maintain performance. A microfluidic system designed to be an integral part of a microchip demonstrates exceptional cooling performance. **See p.211**

An energy-efficient way to improve the performance of electronics systems would be to integrate microfluidic cooling channels into chips, to prevent overheating. However, state-of-the-art microfluidic cooling systems have previously been designed and constructed separately from electronic chips, preventing the channels from being integrated into circuits to provide direct cooling at hotspots. Because such integration greatly increases the complexity of chip fabrication, it would potentially increase the cost. On page 211, van Erp *et al.*[1] report an electronic device designed to have an integrated microfluidic cooling system that closely aligns with the electronic components, and which is constructed using a single, low-cost process.

Power electronics are solid-state electronic devices that convert electrical power into different forms, and are used in a vast array of daily applications[2] — from computers to battery chargers, air conditioners to hybrid electric vehicles, and even satellites. The rising demand for increasingly efficient and smaller power electronics means that the amount of power converted per unit volume of these devices has increased dramatically. This, in turn, has increased the heat flux of the devices — the amount of heat produced per unit area. The heat generated in this way is becoming a big problem: data centres in the United States consume the same amount of energy and water to cool their computer technology as does the city of Philadelphia for its residential needs[1].

Microfluidic cooling systems have great potential for lowering the temperature of electronic devices, because of the efficiency with which heat can be transferred to these systems. In general, three microfluidic cooling designs have been developed. The first is used to cool chips that are covered by a protective lid. Heat is transferred from the chip, through the lid, to a cold plate that contains microfluidic channels through which a liquid coolant flows[3]. Two layers of a thermal interface material (TIM) are used to aid the transfer of heat from the lid to the cold plate: one between the lid and the plate, and the other between the lid and the die (the wafer of semiconductor from which the chip is made).

In the second design, the chip has no lid, and so heat is transferred directly from the back of the chip through a single TIM layer to a microfluidically cooled plate[3]. The main drawback of these two approaches is the need for TIM layers — even though TIMs are designed to transfer heat effectively[4], resistance to heat flow still arises at the interfaces between the TIM layers and the die, lid and cold plate.

An efficient way to overcome this problem is to bring the coolant into direct contact with the chip — this is the third general design. For example, bare-die direct jet cooling is a valuable technique in which a liquid coolant is ejected from nozzles in microchannels directly onto the back of the chip[5–7]. This approach cools highly efficiently because there is no TIM layer, and no changes are needed in the process used to make the chip. However, manufacturing the microfluidics device is generally expensive. Low-cost, polymer-based techniques[8] have been developed, but are not compatible with the existing production and assembly processes for electronic devices.

Another approach that brings coolant into direct contact with the back of the chip is embedded liquid cooling[9,10], in which a cold liquid is pumped through straight, parallel microchannels (SPMCs) etched directly in the semiconductor device. This effectively turns the back of the chip into a heat sink, and offers great cooling performance. However, the die needs extra processing, compared with the other methods. A major drawback of SPMCs is that the pressure in the channels rises considerably as the fluid passes through, which means that a high-power pump is needed. This increases energy consumption and costs, and generates potentially damaging mechanical stress on the semiconductor device. Another big disadvantage is that a high temperature gradient is produced across the chip, which can induce thermo-mechanical stress and

cause local warping of the thin die.

Three-dimensional cooling systems known as embedded manifold microchannels[11,12] (EMMCs) have great potential for reducing pumping-power requirements and temperature gradients compared with SPMCs. In these systems, a 3D hierarchical manifold — a channel component that has several ports for distributing coolant — provides multiple inlets and outlets for embedded microchannels, thereby separating the coolant flow into multiple parallel sections. However, integrating EMMCs into the chips of power electronic devices increases the complexity and cost of constructing the devices. Previously reported EMMCs have therefore been designed and fabricated as separate modules, which are subsequently bonded to a heat source or a commercial chip to assess their cooling properties.

Van Erp *et al.* have made a breakthrough by developing what they describe as a monolithically integrated manifold microchannel (mMMC) — a system in which EMMCs are integrated and co-fabricated with a chip in a single die. The buried channels are therefore embedded right below the active areas of the chip, so that the coolant passes directly beneath the heat sources (Fig. 1).

The construction process for mMMCs involves three steps. First, narrow slits are etched into a silicon substrate coated with a layer of the semiconductor gallium nitride (GaN); the depth of the slits defines the depths of the channels that will be produced. Next, a process known as isotropic gas etching is used to widen the slits in the silicon to the final widths of the channels; this etching process also results in short sections of channels becoming connected to produce longer channel systems. Finally, the openings in the GaN layer at the top of the channels are sealed off with copper. An electronic device can then be fabricated in the GaN layer. Unlike previously reported methods for making manifold microchannels, van Erp and colleagues' process requires no bonding or interfaces between the manifold and devices.

The authors also implemented their design and construction strategy to create a power electronic module that converts alternating current (a.c.) to direct current (d.c.). Experiments with this device show that heat fluxes exceeding 1.7 kilowatts per square centimetre can be cooled using only 0.57 W cm^{-2} of pumping power. Moreover, the liquid-cooled device exhibits significantly higher conversion efficiency than does an analogous uncooled device, because degradation caused by self-heating is eliminated.

Van Erp and colleagues' results are impressive, but as with any technological advance, there is more to be done. For example, the structural integrity of the thin GaN layer needs to be studied over time, to see how long it is stable for. Moreover, the authors used an adhesive that has a maximum operating temperature of 120 °C to connect the microchannels in the devices to fluid-delivery channels in the supporting circuit board. This means that the assembled system would not survive higher temperatures, such as the typical temperature (250 °C) involved during reflow soldering — a process commonly used in the manufacture of electronic devices[13]. Therefore, fluidic connections that are compatible with the temperatures used in manufacturing will need to be developed.

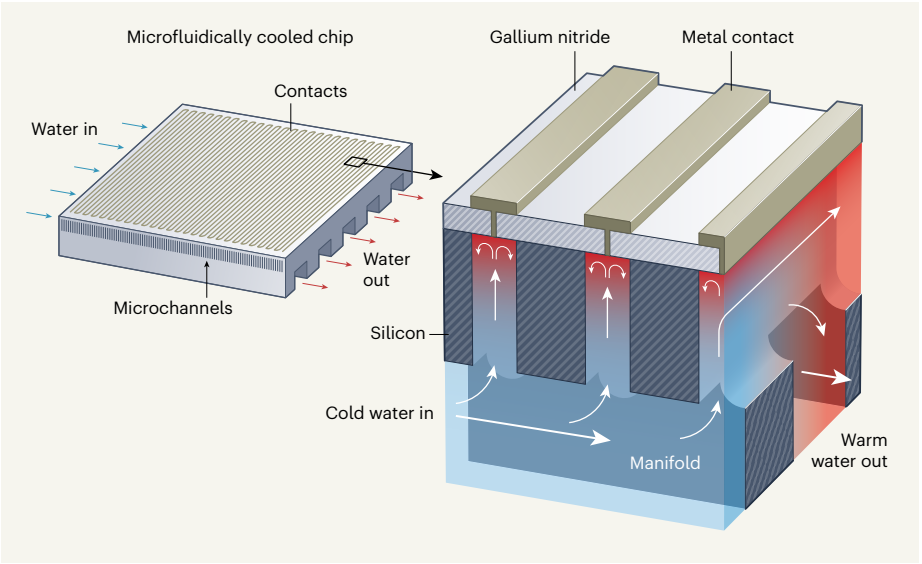


Figure 1 | An integral cooling system for microchips. Van Erp *et al.*[1] have developed a general design for the chips of electronic devices in which a system of microchannels is co-fabricated with the chip, and acts as a cooling system. Cold water is passed through a manifold, which feeds the water into microchannels made of silicon. The water passes directly beneath a layer of gallium nitride, a semiconductor, which contains the components of the electronic device (not shown). The cold water thus efficiently dissipates heat produced by the device, ensuring good performance. Metal contacts at the top seal the channels. (Adapted from Fig. 1a of ref. 1.)

"The authors' work is a big step towards low-cost, ultra-compact and energy-efficient cooling systems for power electronics."

Another future direction of research would be to implement the mMMC concept in a state-of-the-art design for an a.c.-to-d.c. converter — the design reported by van Erp and co-workers is a simple test case. Furthermore, the authors implemented only single-phase cooling with liquid water in their experiments (that is, the water did not get so hot that it became a gas). It would be useful to characterize the cooling and electrical performance of their devices in a two-phase flow-cooling system, in which heat is dissipated by the evaporation of a fluid. Finally, water might not be the ideal coolant for real-world applications, because of the risk of it freezing or coming into direct contact with the chip. Future work should examine the use of different liquid coolants.

Despite the challenges still to be addressed, van Erp and colleagues' work is a big step towards low-cost, ultra-compact and energy-efficient cooling systems for power electronics. Their method outperforms state-of-the -art cooling techniques, and might enable devices that produce high heat fluxes to become part of our daily lives.

Tiwei Wei is at the Interuniversity Microelectronics Centre, Leuven 3001, Belgium, and at KU Leuven, Leuven.
e-mail: tiwei32@stanford.edu

1. van Erp, R., Soleimanzadeh, R., Nela, L., Kampitsis, G. & Matioli, E. *Nature* **585**, 211–216 (2020).
2. Bose, B. K. *IEEE Ind. Electron. Mag.* **3**, 7–11 (2009).
3. Moore, A. L. & Shi, L. *Mater. Today* **17**, 163–174 (2014).
4. Hansson, J., Nilsson, T. M. J., Ye, L. & Liu, J. *Int. Mater. Rev.* **63**, 22–45 (2018).
5. Brunschwiler, T. *et al. Proc. ITherm* 196–203 (2006).
6. Wang, E. N. *et al. J. Microelectromech. Syst.* **13**, 833–842 (2004).
7. Acikalin, T. & Schroeder, C. *Proc. ITherm* 673–679 (2014).
8. Wei, T. *et al. IEEE Int. Electron Devices Meet.* 32.5.1–32.5.4 (2017).
9. Tuckerman, D. B. & Pease, R. F. W. *IEEE Electron Device Lett.* **2**, 126–129 (1981).
10. Phillips, R. J. *Lincoln Lab. J.* **1**, 31–48 (1988).
11. Harpole, G. M. & Eninger, J. E. *Proc. 7th IEEE Semi-Therm Symp.* 59–63 (1991).
12. Jung, K. W. *et al. Proc. ITherm* 98–104 (2017).
13. Pan, J., Toleno, B. J., Chou, T. & Dee, W. J. *Solder. Surf. Mount Technol.* **18**, 48–56 (2006).

Virology

Deep-sleeping HIV genomes under control

Nicolas Chomont

In a few people living with HIV, the virus remains under control without antiretroviral therapy. It emerges that, in these people, the viral DNA that is integrated into the host genome is in a deeply transcriptionally repressed state. **See p.261**

Our ability to keep HIV under control has been revolutionized by antiretroviral therapy (ART). But ART is not a cure — the HIV genome can integrate into host DNA and hide out in cells in a silent form, even after decades of successful therapy[1–3]. ART must be continued throughout life, to prevent the virus from rebounding from these viral reservoirs. Could ways to prevent this viral rebound be found by studying the small proportion (less than 0.5%) of people living with HIV who can control viral replication without the need for ART? On page 261, Jiang *et al.*[4] compared the viral reservoirs of these individuals, known as elite controllers, with those of people who are prescribed ART. Their findings suggest that elite control is associated with a small reservoir from which HIV is unlikely to be reactivated.

The authors began by using a sophisticated sequencing technique to compare viral genomes (proviruses) in millions of cells from the two groups of people. As expected, the comparison revealed fewer copies of the HIV genome in elite controllers than in people receiving ART. However, a higher proportion of the proviruses found in controllers were genetically intact — meaning that they have the potential to generate infectious viral particles when transcribed.

Jiang *et al.* frequently observed many identical copies of the viral genome in elite controllers. This observation confirms[5] that infected cells have the ability to proliferate in controllers, as they do in people receiving ART[6–8]. Elite controllers are known[9] to mount a potent immune response against HIV-infected cells, and the authors found that the proviral sequences persisting in elite controllers were predicted to generate viral proteins that could be targeted by this response.

How, then, do these proviruses escape the immune response? To answer this question, the authors made use of a recently developed approach[10] to analyse the sites at which viruses have integrated into the host genome, in conjunction with corresponding proviral sequences. The analysis revealed several characteristics that suggest that the proviruses found in elite controllers are in a deeper state of latency (dormancy) than are the proviruses in people treated with ART.

First, proviruses in elite controllers are more likely to be integrated in non-protein-coding regions of the genome. Second, viral genomes from controllers are frequently positioned in, or surrounded by, repetitive stretches of DNA at chromosomal structures called centromeres. The host genome is packaged into a DNA–protein complex called chromatin — at centromeres, this packaging is unusually dense, which strongly represses transcription. Third, a substantial portion of HIV genomes in elite controllers are integrated in genes that encode members of the zinc-finger protein family, at which chromatin notoriously carries many molecular modifications that are associated with transcriptional repression[11].

The authors also performed an analysis of accessible chromatin regions (those at which transcription is possible), which revealed that virus-integration sites in the DNA of elite controllers are located significantly farther from accessible chromatin than those in ART-treated individuals. This result reinforces the idea that the genomes of elite controllers are less likely to actively produce viral transcripts and proteins. Indeed, intact proviruses in elite controllers produced ten times fewer viral transcripts than did HIV genomes from people receiving ART.

Two scenarios could explain the peculiar proviral landscape of elite controllers. First, HIV integration could preferentially occur in particular regions of the genomes in these individuals. Alternatively, the proviruses that integrate into non-coding or transcriptionally repressed regions could be selected over time, with those that are more permissive to viral transcription being eliminated.

Definitively distinguishing between these two possibilities would require researchers to follow elite controllers over a long period of time, which was not within the scope of the current study. However, when Jiang *et al.* infected cells from elite controllers and people receiving ART with HIV *in vitro*, they found no significant difference in the integration patterns between the two, making the first scenario unlikely. The second model is also attractive because of the unusually potent immune responses against HIV-infected cells frequently observed in elite controllers. These responses might gradually eliminate the provirus-containing cells that are more likely to produce viral proteins (Fig. 1). Such selection could, over years, result in a reservoir made entirely of proviruses that are unlikely to be reactivated.

This idea is supported by previous work[12] indicating that the pool of replication-

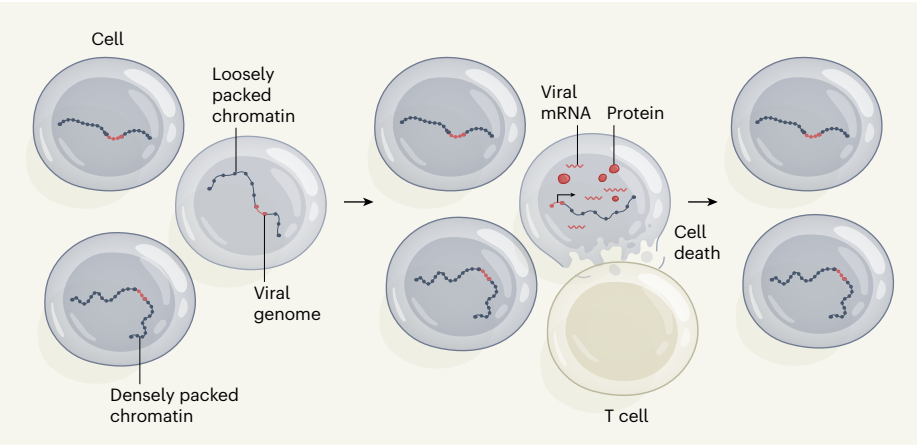


Figure 1 | Selection of sleeping HIV in elite controllers. A small proportion of people living with HIV can control the virus without antiretroviral therapy (ART). Jiang *et al.*[4] provide evidence that the viral DNA in these elite controllers is integrated across the host genome. Some viral genomes become integrated at places in which host DNA is loosely packaged with proteins in a complex called chromatin, meaning that transcription can occur. Other viral DNA is integrated at host sites where transcription is repressed because chromatin packaging is dense. Cells that transcribe the virus (generating viral messenger RNA and proteins) are efficiently targeted by immune T cells — a response seen only in elite controllers. These cells are killed, and so a small pool of cells harbouring deeply latent HIV genomes is evolutionarily selected over time.

competent virus is extremely small in elite controllers. Furthermore, one participant in Jiang and colleagues' study had no detectable replication-competent HIV at all, even though the authors thoroughly analysed more than one billion cells from this person. Whether HIV has been completely eradicated from this individual's body will be hard to demonstrate, but their case is certainly reminiscent of previous reports of HIV cure[13,14].

Elite controllers represent only a small proportion of people living with HIV. Nonetheless, Jiang and colleagues' work has several implications for the rest of this population. It suggests that deeply latent proviruses could preferentially persist after years of viral suppression with ART, particularly in individuals who have maintained immune responses against HIV. Perhaps continuous immune pressure over years would select a small reservoir from which HIV replication would be less likely to reignite. But whether deep-sleeping viral genomes could be reactivated and contribute to viral rebound during ART interruption remains to be determined.

Either way, the results of this study imply that both the intactness and the activation potential of viral genomes should be assessed when measuring the magnitude of the persistent HIV reservoir that can cause viral rebound. Assays that are currently used to estimate the size of the viral reservoir generally measure either the number of intact HIV genomes or their ability to generate RNA or proteins *in vitro*. Jiang and colleagues' work suggests that combining both measures could be necessary, because many intact genomes might not be easily reactivated. A combination measure could provide researchers and clinicians with a better predictor of viral rebound following ART interruption.

The study indicates that a continuous and prolonged cellular-immune pressure might substantially reduce the size of the HIV reservoir over time, by selecting a small pool of cells containing hard-to-reactivate HIV genomes. This, in turn, suggests that immune-cell therapies — including therapies based on CAR T cells, which are currently being developed to control HIV reservoirs[15] — might not only control viral rebound during ART interruption, but also shrink the viral reservoir to a pool of deeply latent proviruses. Whether this could result in a long-term remission of HIV infection remains, of course, to be determined.

Nicolas Chomont is at the CHUM Research Centre, Montreal H2X 0A9, Quebec, Canada, and in the Department of Microbiology, Infectiology and Immunology, Université de Montréal.
e-mail: nicolas.chomont@umontreal.ca

1. Finzi, D. *et al. Science* **278**, 1295–1300 (1997).
2. Wong, J. K. *et al. Science* **278**, 1291–1295 (1997).
3. Chun, T. W. *et al. Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
4. Jiang, C. *et al. Nature* **585**, 261–267 (2020).
5. Boritz, E. A. *et al. Cell* **166**, 1004–1015 (2016).
6. Wagner, T. A. *et al. Science* **345**, 570–573 (2014).
7. Maldarelli, F. *et al. Science* **345**, 179–183 (2014).
8. Simonetti, F. R. *et al. Proc. Natl Acad. Sci. USA* **113**, 1883–1888 (2016).
9. Migueles, S. A. *et al. Immunity* **29**, 1009–1021 (2008).
10. Einkauf, K. B. *J. Clin. Invest.* **129**, 988–998 (2019).
11. Vogel, M. J. *et al. Genome Res.* **16**, 1493–1504 (2006).
12. Blankson, J. N. *et al. J. Virol.* **81**, 2508–2518 (2007).
13. Hutter, G. *et al. N. Engl. J. Med.* **360**, 692–698 (2009).
14. Gupta, R. K. *et al. Nature* **568**, 244–248 (2019).
15. Herzig, E. *et al. Cell* **179**, 880–894.e10 (2019).

This article was published online on 26 August 2020.

Cell biology

The plant response to heat requires phase separation

Simon Alberti

Temperature determines the geographical distribution of plants and their rate of growth and development, but how they sense high temperatures to mount a response was unclear. Now a process underlying this responsiveness is known. **See p.256**

Unlike animals, plants cannot move to escape harsh conditions. Consequently, they must continuously monitor their environment and, when exposed to high temperatures, quickly adjust their expression of developmental and growth-related genes. On page 256, Jung *et al.*[1] describe a molecular process that might underlie this temperature responsiveness.

The expression of developmental and growth-related genes in animals and plants typically occurs in a rhythmic fashion over a 24-hour cycle. Such daily oscillations are controlled by a molecular loop of protein activity that provides what is termed the circadian clock. Clock-induced transcriptional changes enable plants to anticipate daily environmental changes.

In the model plant species *Arabidopsis thaliana*, one component of the circadian clock is a protein assembly called the evening complex. It is maximally active at dusk and represses the expression of many genes important for plant development. The evening complex comprises the transcription-factor protein ELF3 (Fig. 1), a small peptide known as ELF4 and a protein called LUX. Plants with mutations that disable the gene encoding ELF3 flower earlier than normal during development and grow long embryonic stems termed hypocotyls, suggesting that ELF3 has a key developmental role.

Temperature fluctuations are known to affect the circadian rhythm of plants. The growth of *A. thaliana* at 22°C is normally restricted to the period around dawn, because of the repressive action of the evening complex at other times of day[2]. However, at 27°C, this growth repression is relieved[2], and plants show accelerated flowering and rapid hypocotyl elongation compared with growth at 22°C. Yet the mechanism underlying such temperature-regulated growth has remained a mystery. Jung and colleagues propose that a physical process called phase separation is at the heart of plant responsiveness to heat.

To investigate, the authors focused on ELF3. They engineered *A. thaliana* so that ELF3 was replaced with a related version from two plants that do not show temperature-accelerated flowering: *Solanum tuberosum* (potato) and *Brachypodium distachyon* (a grass). The resulting *A. thaliana* plants were indistinguishable from wild-type *A. thaliana* at moderate temperatures, but were unable to accelerate flowering at a higher temperature, suggesting that ELF3 has a key role in temperature responsiveness.

To investigate further, Jung and colleagues focused on a region of ELF3 that is enriched in polar (hydrophilic) amino-acid residues, depleted of charged residues and predicted to be intrinsically disordered. Such protein regions are known as prion-like domains (PrDs), and have been proposed to mediate environmental responses in budding yeast (*Saccharomyces cerevisiae*)[3,4]. Jung *et al.* engineered *A. thaliana* to express a chimaeric protein, in which its normal PrD was replaced with the corresponding region of ELF3 from *B. distachyon*. The authors report that the engineered plant did not display temperature-accelerated flowering, indicating that this ELF3 domain might have a key role in establishing temperature responsiveness.

The PrD of *A. thaliana* contains continuous stretches of the amino acid glutamine that are called polyglutamine (polyQ) repeats. The authors note a correlation between plant species that have long polyQ repeats in this domain and accelerated growth at warm

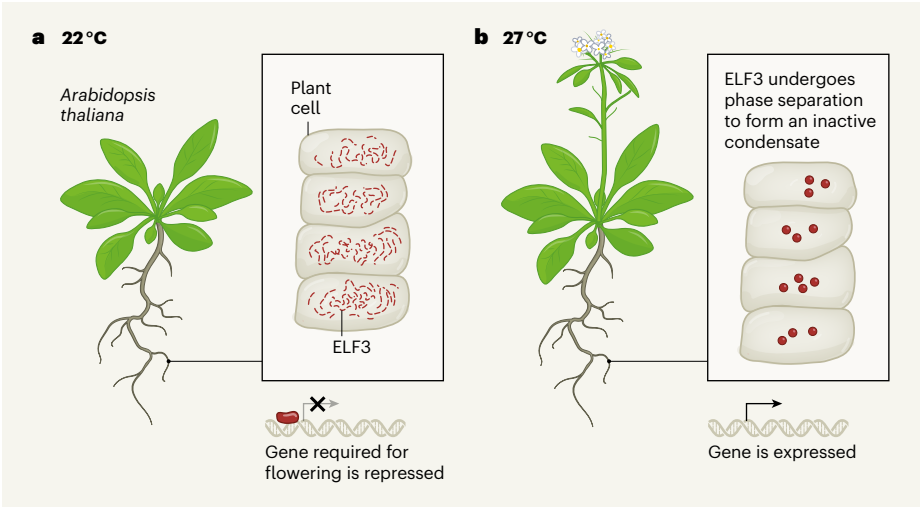


Figure 1 | A mechanism that enables plants to respond to high temperatures. In the model plant *Arabidopsis thaliana*, the protein ELF3 inhibits the expression of certain developmental genes, including some involved in flowering[2]. However, this transcriptional repression is relieved at high temperatures[2]. Jung *et al.*[1] show how this switch in ELF3 activity occurs. **a**, At 22°C, ELF3 is dispersed in the cell in a diffuse pattern, and binds to DNA to block transcription. **b**, At 27°C, ELF3 assembles into 'dots' (also called puncta). The authors suggest that this represents temperature-driven phase separation of ELF3 to form a discrete condensate. This would presumably prevent ELF3 from binding to its target genes, thereby inactivating ELF3 and enabling those genes to be expressed, promoting growth and flowering.

temperatures, suggesting that the polyQ repeats modify ELF3 temperature responsiveness. Because repeat expansions generally evolve rapidly compared with non-repetitive sequences, this correlation suggests a potential way for plants to adapt to the predicted higher temperatures arising from global warming.

To investigate the molecular changes underlying ELF3 temperature responsiveness, the authors used a range of biochemical, biophysical and cell-biological tests. They observed that, at low temperatures, ELF3 was diffusely distributed inside the cell, but when the temperature rose it assembled into microscopically visible 'dots' called puncta. This outcome depended on the presence of the PrD, and the number of observed puncta increased with the length of the polyQ repeats. Crucially, the formation of these puncta was reversed if the temperature fell, suggesting that this represents a normal assembly mechanism in response to heat, rather than an irreversible protein-aggregation event.

Previous work[3–5] led to the proposal that PrD-containing proteins in budding yeast undergo a stimulus-dependent phenomenon called phase separation. This is a process by which a well-mixed protein solution 'demixes' into a dense phase (or condensate) and a dilute phase, comparable to the way that oil and water are partitioned into different phases[6,7]. To test whether the ELF3 PrD forms condensates, the authors performed *in vitro* experiments using a fragment of *A. thaliana* ELF3 containing the PrD. Indeed, this fragment showed temperature-dependent phase separation with a threshold for condensate formation at approximately 28°C. By contrast, the corresponding fragment of ELF3 from *B. distachyon* did not form condensates under the same conditions. This indicates that the *A. thaliana* ELF3 PrD forms condensates *in vitro* in a temperature-dependent manner. However, whether the heat-induced assemblies observed in cells are condensates of inactive ELF3 remains to be established.

Next, the authors focused on ELF4, which binds to ELF3 in the vicinity of its PrD. Jung *et al.* found that ELF4 inhibits the temperature responsiveness of ELF3: plants that were engineered to express higher-than-normal levels of ELF4 were unable to respond to warm temperatures with accelerated flowering. This suggests that the binding of ELF4 to ELF3 modulates condensate assembly by ELF3. The regulation of phase separation by the action of a binding ligand is a widespread phenomenon termed polyphasic linkage[8]. However, more *in vitro* and *in vivo* experiments will be needed to determine whether polyphasic linkage of ELF3 and ELF4 underlies the inhibition of temperature-accelerated flowering.

"A physical process called phase separation is at the heart of plant responsiveness to heat."

The polyQ repeats modify the temperature responsiveness of ELF3, but alone they are probably insufficient to drive this responsiveness, and the identity of the amino-acid residues responsible for driving this property of ELF3 is unknown. Work so far in other systems[9] to understand phase separation of PrD-containing proteins has focused mainly on those that undergo phase separation on cooling, such as a human protein called FUS. Amino-acid residues that are aromatic (those that contain a benzene ring, or an analogue thereof), polar or basic provide cohesive forces for intramolecular interactions in FUS that enable phase separation[9]. ELF3 undergoes phase separation when the temperature rises, rather than falls, and previous studies[5,10] suggest that heat-induced phase separation of elastomeric proteins (flexible proteins with biomechanical functions) often depends on hydrophobic amino-acid residues. Indeed, the ELF3 PrD contains several hydrophobic amino-acid residues, such as methionine, but their role in condensate assembly is unknown.

Another crucial point to establish is how polyQ repeats modify condensate assembly. One possibility is that these repeats alter ELF3 solubility, thus shifting the temperature at which phase separation occurs.

This study raises some exciting questions for the future. For example, what are the properties and composition of these ELF3 condensates in plant cells? Can the ELF3 PrD respond to signals besides temperature, such as other physico-chemical cues? How widespread is this mechanism in plants, and do organisms other than plants regulate components of their circadian clocks through stimulus-dependent phase separation? Repeats of the amino acids threonine and glycine in a transcription-factor protein modulate the temperature responsiveness of the circadian clock in the fruit fly *Drosophila melanogaster*[11]. This suggests that phase separation could have a much broader role in coupling environmental inputs to biological rhythms than had been thought.

Simon Alberti is at the Biotechnology Center, and the Center for Molecular and Cellular Bioengineering, the Technical University of Dresden, Dresden 01307, Germany.
e-mail: simon.alberti@tu-dresden.de

1. Jung, J.-H. *et al. Nature* **585**, 256–260 (2020).
2. Box, M. S. *et al. Curr. Biol.* **25**, 194–199 (2015).
3. Franzmann, T. M. & Alberti, S. *Cold Spring Harb. Perspect. Biol.* **11**, a034058 (2019).
4. Franzmann, T. M. *et al. Science* **359**, eaao5654 (2018).
5. Ruff, K. M., Roberts, S., Chilkoti, A. & Pappu, R. V. *J. Mol. Biol.* **430**, 4619–4635 (2018).
6. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. *Nature Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
7. Shin, Y. & Brangwynne, C. P. *Science* **357**, eaaf4382 (2017).
8. Posey, A. E. *et al. J. Biol. Chem.* **293**, 3734–3746 (2018).
9. Wang, J. *et al. Cell* **174**, 688–699 (2018).
10. Quiroz, F. G. & Chilkoti, A. *Nature Mater.* **14**, 1164–1171 (2015).
11. Huang, Z. J., Curtin, K. D. & Rosbash, M. *Science* **267**, 1169–1172 (1995).

This article was published online on 26 August 2020.

Illuminating the dark spaces of healthcare with ambient intelligence

https://doi.org/10.1038/s41586-020-2669-y

Received: 16 February 2020

Accepted: 14 July 2020

Published online: 9 September 2020

Check for updates

Albert Haque[1], Arnold Milstein[2] & Li Fei-Fei[1,3]✉

Advances in machine learning and contactless sensors have given rise to ambient intelligence—physical spaces that are sensitive and responsive to the presence of humans. Here we review how this technology could improve our understanding of the metaphorically dark, unobserved spaces of healthcare. In hospital spaces, early applications could soon enable more efficient clinical workflows and improved patient safety in intensive care units and operating rooms. In daily living spaces, ambient intelligence could prolong the independence of older individuals and improve the management of individuals with a chronic disease by understanding everyday behaviour. Similar to other technologies, transformation into clinical applications at scale must overcome challenges such as rigorous clinical validation, appropriate data privacy and model transparency. Thoughtful use of this technology would enable us to understand the complex interplay between the physical environment and health-critical human behaviours.

Boosted by innovations in data science and artificial intelligence[1,2], decision-support systems are beginning to help clinicians to correct suboptimal and, in some cases, dangerous diagnostic and treatment decisions[3–5]. By contrast, the translation of better decisions into the physical actions performed by clinicians, patients and families remains largely unassisted[6]. Health-critical activities that occur in physical spaces, including hospitals and private homes, remain obscure. To gain the full dividends of medical advancements requires—in part—that affordable, human-centred approaches are continuously highlighted to assist clinicians in these metaphorically dark spaces.

Despite numerous improvement initiatives, such as surgical safety checklists[7], by the National Institutes of Health (NIH), Centres for Disease Control and Prevention (CDC), World Health Organization (WHO) and private organizations, as many as 400,000 people die every year in the United States owing to lapses and defects in clinical decision-making and physical actions[8]. Similar preventable suffering occurs in other countries, as well-motivated clinicians struggle with the rapidly growing complexity of modern healthcare[9,10]. To avoid overwhelming the cognitive capabilities of clinicians, advances in artificial intelligence hold the promise of assisting clinicians, not only with clinical decisions but also with the physical steps of clinical decisions[6].

Advances in machine learning and low-cost sensors can complement existing clinical decision-support systems by providing a computer-assisted understanding of the physical activities of healthcare. Passive, contactless sensors (Fig. 1) embedded in the environment can form an ambient intelligence that is aware of people's movements and adapt to their continuing health needs[11–14]. Similar to modern driver-assistance systems, this form of ambient intelligence can help clinicians and in-home caregivers to perfect the physical motions that comprise the final steps of modern healthcare. Already enabling better manufacturing, safer autonomous vehicles and smarter sports entertainment[15], clinical physical-action support can more reliably translate the rapid flow of biomedical discoveries into error-free healthcare delivery and worldwide human benefits.

This Review explores how ambient, contactless sensors, in addition to contact-based wearable devices, can illuminate two health-critical environments: hospitals and daily living spaces. With several illustrative clinical-use cases, we review recent algorithmic research and clinical validation studies, citing key patient outcomes and technical challenges. We conclude with a discussion of broader social and ethical considerations including privacy, fairness, transparency and ethics. Additional references can be found in Supplementary Note 1.

Hospital spaces

In 2018, approximately 7.4% of the US population required an overnight hospital stay[16]. In the same year, 17 million admission episodes were reported by the National Health Service (NHS) in the UK[17]. Yet, healthcare workers are often overworked, and hospitals understaffed and resource-limited[18,19]. We discuss a number of hospital spaces in which ambient intelligence may have an important role in improving the quality of healthcare delivery, the productivity of clinicians, and business operations (Fig. 2). These improvements could be of great assistance during healthcare crises, such as pandemics, during which time hospitals encounter a surge of patients[20].

Intensive care units

Intensive care units (ICUs) are specialized hospital departments in which patients with life-threatening illnesses or critical organ failures are treated. In the United States, ICUs cost the health system US$108 billion per year[21] and account for up to 13% of all hospital costs[22].

One promising use case of ambient intelligence in ICUs is the computer-assisted monitoring of patient mobilization. ICU-acquired weaknesses are a common neuromuscular impairment in critically ill

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA. [3]Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA, USA. ✉e-mail: feifeili@stanford.edu

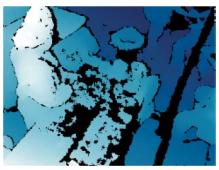
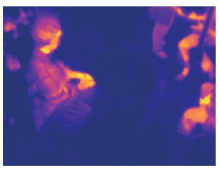
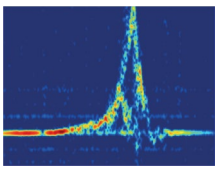
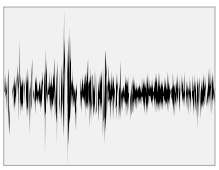
	Camera	Depth sensor	Thermal sensor	Radio sensor	Acoustic sensor
Sensory information	RGB, colour, video	Lidar	Infrared	Radar, Wi-Fi	Microphone
Function	Measures colour (visible light)	Measures distance to objects	Measures surface temperature	Estimates distance and velocity	Measures air pressure waves (sound)
Sampling rate	30 Hz (1,920 × 1,080)	30 Hz (1,280 × 720)	10 Hz (640 × 480)	800 Hz	44.1 kHz
Bit depth	24 bits	16 bits	16 bits	32 bits	16 bits
Uses	Object recognition, person detection	3D object detection, robotic navigation	Night vision, equipment safety	Motion detection, object detection	Speech recognition, event detection
Data visualization					

Fig. 1 | Contactless sensors for ambient intelligence. Brightly coloured pixels denote objects that are closer to the depth sensor. Black pixels denote sensor noise caused by reflective, metallic objects. The radio sensor shows a micro-Doppler signature of a moving object, for which the *x* axis denotes time (5 s) and the *y* axis denotes the Doppler frequency. The radio sensor image is reproduced from ref. [89]. The acoustic sensor displays an audio waveform of a person speaking, for which the *x* axis denotes time (5 s) and the *y* axis denotes the signal amplitude.

patients, potentially leading to a twofold increase in one-year mortality rate and 30% higher hospital costs[23]. Early patient mobilization could reduce the relative incidence of ICU-acquired weaknesses by 40%[24]. Currently, the standard mobility assessment is through direct, in-person observation, although its use is limited by cost impracticality, observer bias and human error[25]. Proper measurement requires a nuanced understanding of patient movements[26]. For example, localized wearable devices can detect pre-ambulation manoeuvres (for example, the transition from sitting to standing)[27], but are unable to detect external assistance or interactions with the physical space (for example, sitting on chair versus bed)[27]. Contactless, ambient sensors could provide the continuous and nuanced understanding needed to accurately measure patient mobility in ICUs.

In one pioneering study, researchers installed ambient sensors (Fig. 2a) in one ICU room (Fig. 2b) and collected 362 h of data from eight patients[28]. A machine-learning algorithm categorized in-bed, out-of-bed and walking activities with an accuracy of 87% when compared to retrospective review by three physicians. In a larger study at a different hospital (Fig. 2c), another research team installed depth sensors in eight ICU rooms[29]. They trained a convolutional neural network[1] on 379 videos to categorize mobility activities into four categories (Fig. 2d). When validated on an out-of-sample dataset of 184 videos, the algorithm demonstrated 87% sensitivity and 89% specificity. Although these preliminary results are promising, a more insightful evaluation could provide stratified results rather than aggregate performance on short, isolated video clips. For example, one study used cameras, microphones and accelerometers to monitor 22 patients in ICUs, with and without delirium, over 7 days[30]. The study found significantly fewer head motions of patients who were delirious compared with patients who were not. Future studies could leverage this technology to detect delirium sooner and provide researchers with a deeper understanding of how patient mobilization affects mortality, length of stay and patient recovery.

Another early application is the control of hospital infections. Worldwide, more than 100 million patients are affected by hospital-acquired (that is, nosocomial) infections each year[31], with up to 30% of patients in ICUs experiencing a nosocomial infection[32]. Proper compliance with hand hygiene protocols is one of the most effective methods of reducing the frequency of nosocomial infections[33]. However, measuring compliance remains challenging. Currently, hospitals rely on auditors to measure compliance, despite being expensive, non-continuous and biased[34]. Wearable devices, particularly radio-frequency identification (RFID) badges, are a potential solution. Unfortunately, RFID provides coarse location estimates (that is, within tens of centimetres[35]), making it unable to categorize fine-grained movements such as the WHO's five moments of hand hygiene[36]. Alternatively, ambient sensors could monitor handwashing activities with higher fidelity—differentiating true use of an alcohol-gel dispenser from a clinician walking near a dispenser. In a pioneering study, researchers installed depth sensors above wall-mounted dispensers across an entire hospital unit[37,38]. A deep-learning algorithm achieved an accuracy of 75% at measuring compliance for 351 handwashing events during one hour. During the same time period, an in-person observer was 63% accurate, while a proximity algorithm (for example, RFID) was only 18% accurate. In more nuanced studies, ambient intelligence detected the use of contact-precautions equipment[39] and physical contact with the patient[40]. A critical next step is to translate ambient observation into changes in clinical behaviour, with a goal of improving patient outcomes.

Operating rooms

Worldwide, more than 230 million surgical procedures are undertaken annually[41] with up to 14% of patients experiencing an adverse event[42]. This percentage could be reduced through quicker surgical feedback, such as more frequent coaching of technical skill, which could reduce the number of errors by 50%[43]. Currently, the skills of a surgeon are assessed by peers and supervisors[44], despite being time-consuming, infrequent and subjective. Wearable sensors can be attached to hands or instruments to estimate the surgeon's skills[45], but may inhibit hand dexterity or introduce sterilization complexity. Ambient cameras are an unobtrusive alternative[46]. One study trained a convolutional neural network[1] to track a needle driver in prostatectomy videos[47]. Using peer-evaluation as the reference standard, the algorithm categorized 12 surgeons into high- and low-skill groups with an accuracy of 92%. A different study used videos from ten cholecystectomy procedures to reconstruct the trajectories of instruments during surgery and linked them to technical ratings by expert surgeons[48]. Further studies, such as video-based surgical phase recognition[49], could potentially lead to improved surgical training. However, additional clinical validation is needed and appropriate feedback mechanisms must be tested.

In the operating room, ambient intelligence is not limited to endoscopic videos[50]. Another example is the surgical count—a process of counting used objects to prevent objects being accidentally retained inside the patient[51]. Currently, dedicated staff time and effort are required to visually and verbally count these objects. Owing to attention deficit and insufficient team communication[52], it is possible for the human-adjudicated count to incorrectly label an object as returned

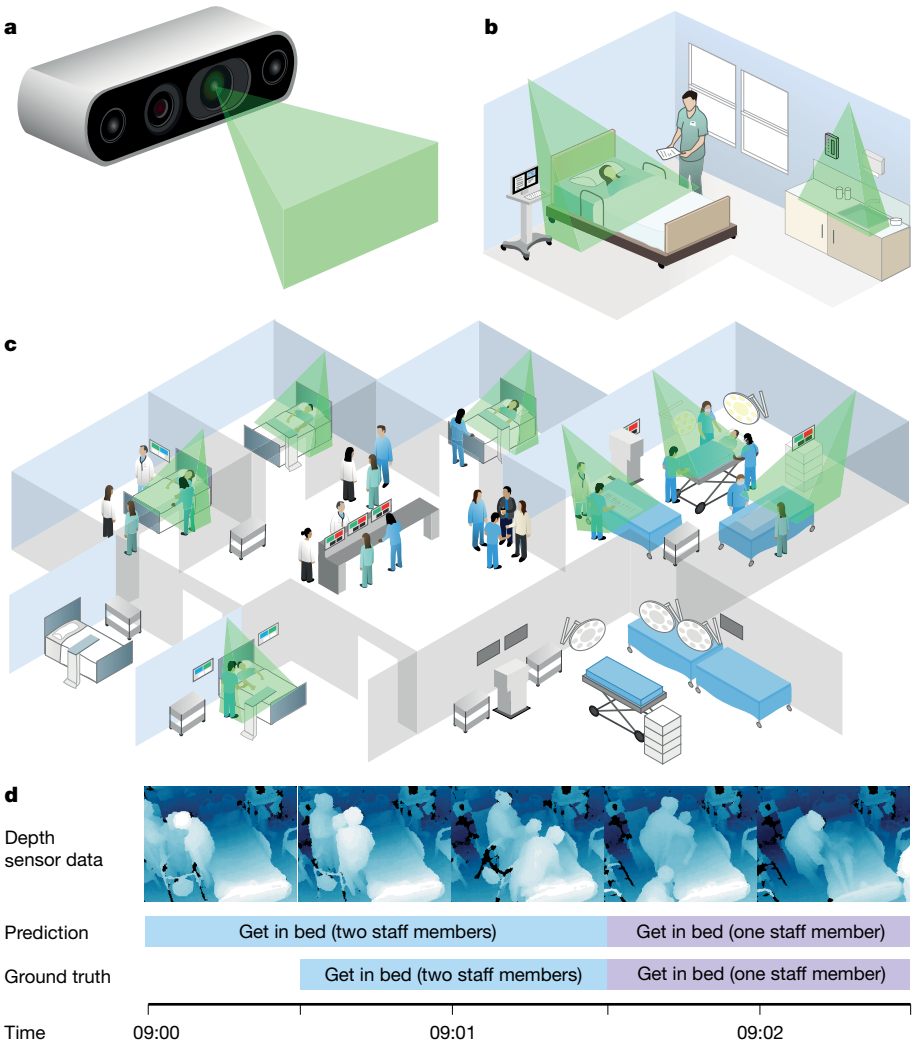


Fig. 2 | Ambient intelligence for hospitals. a, Commercial ambient sensor for which the coverage area is shown in green (that is, the field of view of visual sensors and range for acoustic and radio sensors). **b**, Sensors deployed inside a patient room can capture conversations and the physical motions of patients, clinicians and visitors. **c**, Sensors can be deployed throughout a hospital. **d**, Comparison of predictions and ground truth of activity from depth sensor data. Top, data from a depth sensor. Middle, the prediction of the algorithm of mobilization activity, duration and the number of staff who assist the patient. Bottom, human-annotated ground truth from a retrospective video review. **d**, Adapted from ref. [29].

when it is actually missing[51]. Automated counting systems, in particular, could assist surgical teams[53]. One study showed that barcode-equipped laparotomy sponges reduced the retained object rate from once every 16 days to once every 69 days[54]. Similar results were found with RFID and Raytec sponges[55]. However, owing to their size, barcodes and RFID cannot be applied to needles and instruments, which are responsible for up to 55% of counting discrepancies[51]—each discrepancy delaying the case by 13 min on average[51]. In addition to sponges, ambient cameras could count these smaller objects and potentially staff members[56]. In one operating room, researchers used ceiling-mounted cameras to track body parts of surgical team members with errors as low as five centimetres[57]. Ambient data collected throughout the room could create fine-grained logs of intraoperative activity[58]. Although these studies are promising as a proof of concept, further research needs to quantify the impact on patient outcomes, reimbursement and efficiency gains.

Other healthcare spaces

Clinicians spend up to 35% of their time on medical documentation tasks[59], taking valuable time away from patients. Currently, physicians perform documentation during or after each patient visit. Some providers use medical scribes to alleviate this burden, resulting in 0.17 more patients seen per hour and 0.21 more relative value units per patient (that is, insurer reimbursement)[60]. However, scribes are expensive to train and have high turnover[61]. Ambient microphones could perform a similar task to that of medical scribes[62]. Medical dictation software is an alternative, but is traditionally limited to the post-visit report[63]. In one study, researchers trained a deep-learning model on 14,000 h of outpatient audio from 90,000 conversations between patients and physicians[64]. The model demonstrated a word-level transcription accuracy of 80%, suggesting it may be better than the 76% accuracy of medical scribes[65]. In terms of clinical utility, one medical provider found that microphones attached to eyeglasses reduced time spent on documentation from 2 h to 15 min and doubled the time spent with patients[62].

From a management standpoint, ambient intelligence can improve the transition to activity-based costing[66]. Traditionally, insurance companies and hospital administrators estimated health outcomes per US dollar spent through a top-down approach of value-based accounting[67]. Time-driven activity-based costing is a bottom-up alternative and estimates the costs by individual resource time and cost (for example, the use of an ICU ventilator for 48 h)[68]. This can better inform process redesigns[66]—which, for one provider, led to 19% more patient visits with 17% fewer employees, without degradation of the patient outcomes[69]. Currently, in-person observations, staff interviews and electronic health records are used to map clinical activities to costs[68]. As described in

this Review, ambient intelligence can automatically recognize clinical activities[70], count healthcare personnel[29] and estimate the duration of activities[29] (Fig. 2d). However, evidence of the clinical benefits of ambient intelligence is currently lacking, as the paradigm of activity-based costing is relatively new to hospital staff. As the technology develops, we hope that hospital administrators participate in the implementation and validation of ambient activity-based costing systems.

Daily living spaces

Humans spend a considerable portion of time at home. Around the world, the population is ageing[71]. Not only will this increase the amount of time spent at home, but it will also increase the importance of independent living, chronic disease management, physical rehabilitation and mental health of older individuals in daily living spaces.

Elderly living spaces and ageing

By 2050, the world's population aged 65 years or older will increase from 700 million to 1.5 billion[71]. Activities of daily living (ADLs), such as bathing, dressing and eating, are critical to the well-being and independence of this population. Impairment of one's ability to perform ADLs is associated with a twofold increase in falling risk[72] and up to a fivefold increase in one-year mortality rate[73]. Earlier detection of impairments could provide an opportunity to provide timely clinical care[11], potentially improving the ability to perform ADL by a factor of two[74]. Currently, ADLs are measured through self-reported questionnaires or manual grading by caregivers, despite the fact that these measurements are infrequent, biased and subjective[75]. Alternatively, wearable devices (such as accelerometers or electrocardiogram sensors) can track not only ADLs, but also heart rate, glucose level and respiration rate[76]. However, wearable devices are unable to discern whether a patient received ADL assistance—a key component of ADL evaluations[77]. Contactless, ambient sensors (Fig. 3a) could potentially identify these clinical nuances while detecting a greater range of activities[78].

In one of the first studies of its kind, researchers installed a depth and thermal sensor (Fig. 3b) inside the bedroom of an older individual and observed 1,690 activities during 1 month, including 231 instances of caregiver assistance[79] (Fig. 3c). A convolutional neural network[1] was 86% accurate at detecting assistance. In a different study, researchers collected ten days of video from six individuals in an elderly home and achieved similar results[80]. Although visual sensors are promising, they raise privacy concerns in some environments, such as bathrooms, which is where grooming, bathing and toileting activities occur, all of which are strongly indicative of cognitive function[81]. This led researchers to explore acoustic[82] and radar sensors[83]. One study used microphones to detect showering and toileting activities with accuracies of 93% and 91%, respectively[82]. However, a limitation of these studies is their evaluation in a small number of environments. Daily living spaces are highly variable, thus introducing generalization challenges. Additionally, privacy is of utmost importance. Development and verification of secure, privacy-safe systems is essential if this technology is to illuminate daily living spaces.

Another application for the independent living of older individuals is fall detection[84]. Approximately 29% of community-dwelling adults fall at least once a year[85]. Laying on the floor for more than one hour after a fall is correlated with a fivefold increase in 12-month mortality[86]. Furthermore, the fear of falling—associated with depression and lower quality of life[87]—can be reduced due to the perceived safety benefit of fall-detection systems[88]. For decades, researchers developed fall-detection systems with wearable devices and contactless ambient sensors[89]. A systematic review found that wearable devices detected falls with 96% accuracy while ambient sensors were 97% accurate[90]. In a different study, researchers installed Bluetooth (that is, radio) beacons in 271 homes[91]. Using signal strengths from each beacon, a machine-learning algorithm categorized the frailty of older individuals with an accuracy of 98%. In another study, researchers installed depth and radar sensors on the ceiling of 16 senior living apartments for 2 years[92]. Radar signals, transformed by a wavelet decomposition, detected 100% of falls with fewer than two false alarms per day[93]. Depth sensors produced one false alarm per month with a fall detection rate of 98%[94]. The ambient sensors were sufficiently fast (that is, low latency) to provide real-time email alerts to caregivers at 13 assisted-living communities[95]. Compared to a control group of 85 older individuals over 1 year, the real-time intervention significantly slowed the functional decline of 86 older individuals. When combined with wearable devices, one study found that the fall-detection accuracy of depth sensors increased from 90% to 98%[96], suggesting potential synergies between contactless and wearable sensors. As ambient intelligence begins to bridge the gap between observation and intervention, further studies are needed to explore regulatory approval processes, legal implications and ethical considerations.

Chronic disease management

With applications to physical rehabilitation and chronic diseases, gait analysis is an important tool for diagnostic testing and measuring treatment efficacy[97]. For example, frequent and accurate gait analysis could improve postoperative health for children with cerebral palsy[98] or enable earlier detection of Parkinson's disease by up to 4 years[99]. Traditionally limited to research laboratories with force plates and motion capture systems[100], gait analysis is being increasingly conducted with wearable devices[101]. One study used accelerometers to estimate the clinical-standard 6-min walking distance of 30 patients with chronic lung disease[102]. The study found an average absolute error rate of 6%. One limitation is that wearables must be physically attached to the body, making them inconvenient for patients[103]. Alternatively, contactless sensors could continuously measure gait with improved fidelity and create interactive, home-based rehabilitation programmes[104]. Several studies measured gait in natural settings with cameras[105], depth sensors[104], radar[106] and microphones[107]. One study used depth sensors to measure gait patterns of nine patients with Parkinson's disease[108]. Using a high-end motion capture system as the ground truth, the study found that depth sensors could track vertical knee motions to within four centimetres. Another study used depth sensors to create an exercise game for patients with cerebral palsy[109]. Over the course of 24 weeks, patients using the game improved their balance and gait by 18% according to the Tinetti test[110]. Although promising, these studies evaluated a single sensor modality. In laboratory experiments, gait detection improved by 3% to 7% when microphones were combined with wearable sensors[111]. When feasible, studies could investigate potential synergies of multiple sensing modalities (such as passive infrared motion sensors, contact sensors and wearable cameras).

Mental health

Mental illnesses, such as depression, anxiety and bipolar disorder, affect 43 million adults in the USA[112] and 165 million people in the European Union[113]. It is estimated that 56% of adults with mental illnesses do not seek treatment owing to barriers such as financial cost and provider availability[112]. Currently, self-reported questionnaires and clinical evaluations (for example, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)) are the standard tool for identifying symptoms of mental illness, despite being infrequent and biased[114]. Alternatively, ambient sensors could provide continuous and cost-effective symptom screening[115]. In one study, researchers collected audio, video and depth data from 69 individuals during 30-min, semi-structured clinical interviews[116]. Using the patient's verbal cues and upper body movement, a machine-learning algorithm detected 46 patients with schizophrenia with a positive predictive value of 95% and sensitivity of 84%. Similarly, in an emergency department, natural language analysis of clinical interviews with 61 adolescent individuals, of whom 31 were suicidal, yielded a model capable of categorizing patients who were suicidal with 90%

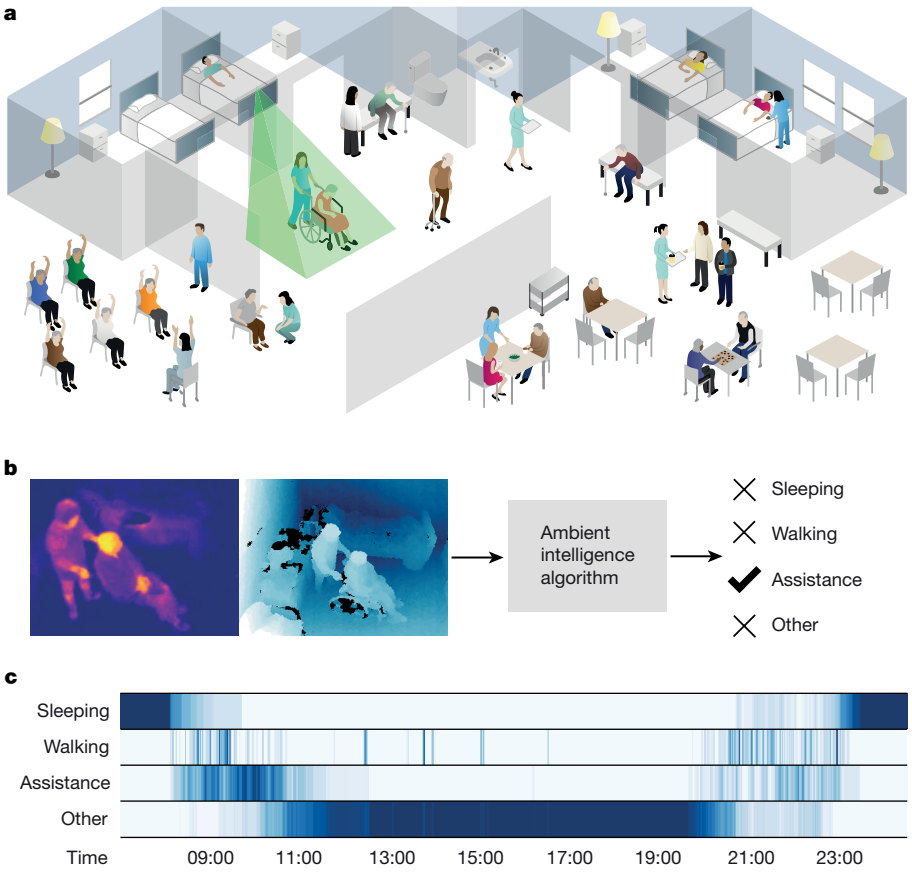


Fig. 3 | Ambient intelligence for daily living spaces. a, Elderly home equipped with one ambient sensor. The green frustum indicates the coverage area of the sensor (that is, the field of view for visual sensors and range for acoustic and radio sensors). **b**, Thermal and depth data from the sensor are processed by an ambient intelligence algorithm for activity categorization. **c**, Summary of a patient's activities for a single day. Darker blue sections indicate more frequent activity. **c**, Adapted from ref. [79].

accuracy[117]. Although impressive, further trials are needed to validate the effect on patient outcomes.

However, even after detection, treating mental health illnesses remains complex. Idiosyncratic therapist effects can cause up to 17% of the variance in outcomes[118], making it difficult to conduct psychotherapy research. Transcripts are the standard method for identifying features of good therapy[119], but are expensive to collect. Manual coding of a 20-min session can range from 85 to 120 min[120]. Ambient sensors could provide cheaper, higher-quality transcripts for psychotherapy research. Using text messages of treatment sessions, one study used a recurrent neural network[1] to detect instances of 24 therapist techniques from 14,899 patients[121]. The study identified several techniques correlated with improved Patient Health Questionnaire (PHQ-9) and General Anxiety Disorder (GAD-7) scores. A different study used microphones and a speech-recognition algorithm to transcribe and estimate therapists' empathy from 200 twenty-minute motivational interviewing sessions[120]. Using a committee of human assessors as the gold standard, the algorithm was 82% accurate. Although this is lower than the 90% accuracy of a single human assessor[120], ambient intelligence can more readily be applied to a larger number of patients. Using ambient intelligence, researchers can now conduct large-scale studies to reaffirm their understanding of psychotherapy frameworks. However, further research is needed to validate the generalization of these systems to a diverse population of therapists and patients.

Technical challenges and opportunities

Ambient intelligence can potentially illuminate the healthcare delivery process by observing recovery-related behaviours, reducing unintended clinician errors, assisting the ageing population and monitoring patients with chronic diseases. In Table 1, we highlight seven technical challenges and opportunities related to the recognition of human behaviour in complex scenes and learning from big data and rare events in clinical settings.

Behaviour recognition in complex scenes

Understanding complex human behaviours in healthcare spaces requires research that spans multiple areas of machine intelligence such as visual tracking, human pose estimation and human–object interaction models. Consider morning rounds in a hospital. Up to a dozen clinicians systematically review and visit each patient in a hospital unit. During this period, clinicians may occlude a sensor's view of the patient, potentially allowing health-critical activities to go undetected. If an object is moving before occlusion, tracking algorithms (Table 1) can estimate the position of the object while occluded[122]. For longer occlusions, matrix completion methods, such as image inpainting, can 'fill in' what is behind the occluding object[123]. Similar techniques can be used to denoise audio in spectrogram form[124]. If there are no occlusions, the next step is to locate people. During morning rounds, clinicians may hand each other objects or point across the room, introducing multiple layers of body parts from the perspective of the sensor. Human pose-estimation algorithms (Table 1) attempt to resolve this ambiguity by precisely locating body parts and assigning them to the correct individuals[125]. Building highly accurate human behaviour models is needed for ambient intelligence to succeed in complex clinical environments.

Ambient intelligence needs to understand how humans interact with objects and other people. One class of methods attempts to identify visually grounded relationships in images[126], commonly in the form of

Table 1 | Algorithmic challenges

Challenge	Sub-challenge	Technical approaches	ICUs		Operating rooms		Other		Elderly care		Chronic	Mental health	
			Patient mobility	Hand hygiene	Skills	Surgical count	Notes	Costing	ADLs	Falls	Gait analysis	Symptom screening	Therapy research
Behaviour recognition in complex scenes	Complex environments	Visual tracking, matrix completion	x	x	x	x	–	x	x	x	x	x	–
	Locating multiple humans	Pedestrian detection, human pose estimation	x	–	–	–	–	x	x	x	x	–	–
	Recognizing human behaviours	Scene graphs, activity recognition	x	x	–	x	–	x	x	x	–	–	x
Learning with big data and rare events	Big data	Distributed learning, optimizers	x	x	x	x	x	x	x	x	x	x	x
	Real-time detections	Two-stage models, model compression	–	x	x	x	–	–	–	x	–	–	–
	Rare events	Calibration, loss weighting	x	–	–	x	–	–	–	x	–	x	–
	Generalization to new environments	Transfer learning, few-shot learning	x	x	x	x	x	x	x	x	x	x	x

Rows denote algorithmic challenges. Columns denote clinical-use cases. Challenges applicable to specific clinical-use cases are marked by an 'x'. 'Skills' indicates the evaluation of surgical skills; 'notes' refers to medical documentation.

a scene graph (Table 1). A scene graph is a network of interconnected nodes, in which each node represents an object in the image and each connection represents their relationship[127]. Not only can scene graphs aid in the recognition of human behaviour, but they could also make ambient intelligence more transparent[128].

Learning from big data and rare events

Ambient sensors will produce petabytes of data from hospitals and homes[129]. This requires new machine-learning methods that are capable of modelling rare events and handling big data to be developed (Table 1). Large-scale activity-understanding models could require days to train unless large clusters of specialized hardware are used[130]. Cloud servers are a potential solution, but can be expensive as ambient intelligence may require considerable storage, computation and network bandwidth. Improved gradient-based optimizers[131] and neural network architectures[132] can potentially reduce training time. However, quickly training a model does not guarantee it will be fast during inference (that is, real-time detections) (Table 1). For example, video-based activity recognition models are slow, typically on the order of 1 to 10 frames per second[133]. Even optimized models capable of 100 frames per second[134] may have difficulties processing terabytes of data each day. Techniques such as model compression[135] and quantization[136] can reduce storage and computational requirements. Instead of processing audio or video at full spatial or temporal resolution, some methods quickly identify segments of interest, known as proposals[137]. These proposals are then provided to heavy-duty modules for highly accurate but computation-intensive activity recognition.

Although the volume of data produced by ambient sensors is large, some clinical events are rare and infrequent (Table 1). The detection of these long-tail events is necessary to understand health-critical behaviours. Consider the example of fall detection. The majority of ambient data contains normal activity, biasing the algorithm owing to label imbalance. More broadly, statistical bias can apply to any category of data, such as protected class attributes[138]. One solution is to statistically calibrate the algorithm, resulting in consistent error rates across specified attributes[139]. However, some healthcare environments may have a greater incidence of falls than in the original training set. This requires generalization (Table 1): the ability of an algorithm to operate on unseen distributions[140]. Instead of training a model designed for all distributions, one alternative is to take an existing model and fine-tune it on the new distribution[141]—also known as transfer learning[142]. Another solution, domain adaptation[143], attempts to minimize the gap between the training and testing distributions, often through better feature representations. For low-resource healthcare providers, few-shot learning—algorithms capable of learning from as few as one or two examples[144]—could be used.

Social and ethical considerations

Trustworthiness of ambient intelligence systems is critical to achieve the potential of this technology. Although there is an increasing body of literature on trustworthy artificial intelligence[145], we consider four separate dimensions of trustworthiness: privacy, fairness, transparency and research ethics. Developing the technology while addressing all four factors requires close collaborations between experts from medicine, computer science, law, ethics and public policy.

Privacy

Ambient sensors, by design, continuously observe the environment and can uncover new information about how physical human behaviours influence the delivery of healthcare. For example, sensors can measure vital signs from a distance[146]. While convenient, such knowledge could potentially be used to infer private medical conditions. As citizens worldwide are becoming more sensitive to mass data collection, there are growing concerns over confidentiality, sharing and retention of this information[147]. It is therefore essential to co-develop this technology with privacy and security in mind, not only in terms of the technology itself but also in terms of a continuous involvement of all stakeholders during the development[148].

A number of existing and emerging privacy-preserving techniques are presented in Fig. 4. One method is to de-identify data by removing the identities of the individuals. Another method is data minimization,

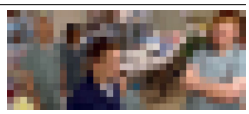
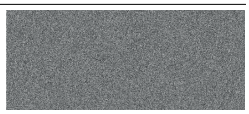
Method	Description	Computing hardware	Transformed result
Differential privacy	Adds noise to the data; minimally affects population-level analysis	Edge computer	
Face blurring	Detects and blurs human faces	Sensor, edge computer	
Dimensionality reduction	Reduces the input size by reducing the number of features	Sensor, edge computer	
Body masking	Replaces peoplewith faceless avatars	Edge computer	
Federated learning	Edge devices learn locally, then sends gradient updates to central server	Edge computer, centralized server	
Homomorphic encryption	Enables predictions to be made from encrypted data	Edge computer, centralized server	

Fig. 4 | Computational methods to protect privacy. There is a trade-off between the level of privacy protection provided by each method and the required computational resources. The methods used to generate the transformed images are described in detail elsewhere: differential privacy, ref. [166]; dimensionality reduction, ref. [167]; body masking, ref. [168]; federated learning, ref. [169]; homomorphic encryption, ref. [170]. The original image was produced by S. McCoy and has previously been published[171]. The appearance of US Department of Defence visual information does not imply or constitute endorsement by the US Department of Defence.

which minimizes data capture, transport and human bycatch. An ambient system could pause when a hospital room is unoccupied by a patient. However, even if data are de-identified, it may be possible to re-identify an individual[149]. Super-resolution techniques[150] can partially reverse the effects of face blurring and dimensionality reduction techniques, potentially enabling re-identification. This suggests that data should remain on-device to reduce the risk of unauthorized access and re-identification.

Legal and social complexities will inevitably arise. There are documented examples in which companies were required to provide data from ambient speakers and cameras to law enforcement[151]. Although these devices were located inside potential crime scenes, this raises the question at what point incidental findings outside the crime scene, such as inadvertent confessions, should be disclosed. Related to data sharing, some healthcare organizations have shared patient information with third parties such as data brokers[152]. To mitigate this, patients should proactively request healthcare providers to use privacy-preserving practices (Fig. 4). Additionally, clinicians and technologists must collaborate with critical stakeholders (for example, patients, family or caregivers), legal experts and policymakers to develop governance frameworks for ambient systems.

Fairness

Ambient intelligence will interact with large patient populations, potentially several orders of magnitude larger than the reach of current clinicians. This compels us to scrutinize the fairness of ambient systems. Fairness is a complex and multi-faceted topic, discussed by multiple research communities[138]. We highlight here two aspects of algorithmic fairness as examples: dataset bias and model performance.

Labelled datasets are the foundation of most machine-learning systems[1]. However, medical datasets have been biased, even before deep learning[153]. These biases can adversely affect clinical outcomes for certain populations[154]. If an individual is missing specific attributes, whether owing to data-collection constraints or societal factors, algorithms could misinterpret their entire record, resulting in higher levels of predictive error[155]. One method for identifying bias is to analyse model performance across different groups[156]. In one study, error rates varied across ethnic groups when predicting 30-day psychiatric readmission rates[157]. A more rigorous method could test for equal sensitivity and equal positive-predictive value. However, equal model performance may not produce equal clinical outcomes, as some populations may have inherent physiological differences. Nonetheless, progress is being made to mitigate bias, such as the PROBAST tool[158].

Transparency

Ambient intelligence can uncover insights about how healthcare delivery is influenced by human behaviour. These discoveries may surprise some researchers, in which case, clinicians and patients need to trust the findings before using them. Instead of opaque, black-box models, ambient intelligence systems should provide interpretable results that are predictive, descriptive and relevant[159]. This can aid in the challenging task of acquiring stakeholder buy-in, as technical illiteracy and model opacity can stagnate efforts to use ambient intelligence in healthcare[160]. Transparency is not limited to the algorithm. Dataset transparency—a detailed trace of how a dataset was designed, collected and annotated—would allow for specific precautions to be taken for future applications, such as training human annotators or revising the inclusion and exclusion criteria of a study. Formal guidelines on transparency, such as the TRIPOD statement[161], are actively being developed. Another tool is the use of model cards[162], which are short analyses that benchmark the algorithm across different populations and outline evaluation procedures.

Research ethics

Ethical research encompasses topics such as the protection of human participants, independent review and public beneficence. The Belmont Report, which prompted the regulation of research involving human participants, includes 'respect for persons' as a fundamental principle. In research, this manifests as informed consent from research participants. However, some regulations allow research to occur without consent if the research poses minimal risks to participants or if it is infeasible to obtain consent. For large-scale ambient intelligence studies, obtaining informed consent can be difficult, and it may in some cases be impossible due to automatic de-identification techniques (Fig. 4). In these cases, public engagement or deliberative democracy can be alternative solutions[163].

Relying solely on the integrity of principal investigators to conduct ethical research may introduce potential conflicts of interest. To mitigate this risk, academic research that involves human participants requires the approval from an Institutional Review Board. Public health surveillance, intended to prevent widespread disease and improve health, does not require independent review[164]. Depending on the application, ambient intelligence could be classified as either[165]. Researchers are urged to consult with experts from law and ethics to determine appropriate steps for protecting all human participants while maximizing public beneficence.

Summary

Centuries of medical practice led to a knowledge explosion, fuelling unprecedented advances in human health. Breakthroughs in artificial intelligence and low-cost, contactless sensors have given rise to an ambient intelligence that can potentially improve the physical execution of healthcare delivery. Preliminary results from hospitals and

daily living spaces confirm the richness of information gained through ambient sensing. This extraordinary opportunity to illuminate the dark spaces of healthcare requires computer scientists, clinicians and medical researchers to work closely with experts from law, ethics and public policy to create trustworthy ambient intelligence systems for healthcare.

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
This paper reviews developments in deep learning and explains common neural network architectures such as convolutional and recurrent neural networks when applied to visual and natural language-processing tasks.
2. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
3. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
This perspective describes the use of computer vision, natural language processing, speech recognition and reinforcement learning for medical imaging tasks, electronic health record analysis, robotic-assisted surgery and genomic research.
4. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
This review outlines how artificial intelligence is used by clinicians, patients and health systems to interpret medical images, find workflow efficiencies and promote patient self-care.
5. Sutton, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* **3**, 17 (2020).
6. Yeung, S., Downing, N. L., Fei-Fei, L. & Milstein, A. Bedside computer vision — moving artificial intelligence from driver assistance to patient safety. *N. Engl. J. Med.* **378**, 1271–1273 (2018).
7. Haynes, A. B. et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *N. Engl. J. Med.* **360**, 491–499 (2009).
8. Makary, M. A. & Daniel, M. Medical error—the third leading cause of death in the US. *Br. Med. J.* **353**, i2139 (2016).
9. Tallentire, V. R., Smith, S. E., Skinner, J. & Cameron, H. S. Exploring error in team-based acute care scenarios: an observational study from the United Kingdom. *Acad. Med.* **87**, 792–798 (2012).
10. Yang, T. et al. Evaluation of medical malpractice litigations in China, 2002–2011. *J. Forensic Sci. Med.* **2**, 185–189 (2016).
11. Pol, M. C., ter Riet, G., van Hartingsveldt, M., Kröse, B. & Buurman, B. M. Effectiveness of sensor monitoring in a rehabilitation programme for older patients after hip fracture: a three-arm stepped wedge randomised trial. *Age Ageing* **48**, 650–657 (2019).
12. Fritz, R. L. & Dermody, G. A nurse-driven method for developing artificial intelligence in "smart" homes for aging-in-place. *Nurs. Outlook* **67**, 140–153 (2019).
13. Kaye, J. A. et al. F5-05-04: ecologically valid assessment of life activities: unobtrusive continuous monitoring with sensors. *Alzheimers Dement.* **12**, P374 (2016).
14. Acampora, G., Cook, D. J., Rashidi, P. & Vasilakos, A. V. A survey on ambient intelligence in health care. *Proc IEEE Inst. Electr. Electron. Eng.* **101**, 2470–2494 (2013).
15. Cook, D. J., Duncan, G., Sprint, G. & Fritz, R. Using smart city technology to make healthcare smarter. *Proc IEEE Inst. Electr. Electron. Eng.* **106**, 708–722 (2018).
16. Centers for Disease Control and Prevention. *National Health Interview Survey: Summary Health Statistics* https://www.cdc.gov/nchs/nhis/shs.htm (2018).
17. NHS Digital. *Hospital Admitted Patient Care and Adult Critical Care Activity 2018–19* https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2018-19 (NHS, 2019).
18. Patel, R. S., Bachu, R., Adikey, A., Malik, M. & Shah, M. Factors related to physician burnout and its consequences: a review. *Behav. Sci. (Basel)* **8**, 98 (2018).
19. Lyon, M. et al. Rural ED transfers due to lack of radiology services. *Am. J. Emerg. Med.* **33**, 1630–1634 (2015).
20. Adams, J. G. & Walls, R. M. Supporting the health care workforce during the COVID-19 global epidemic. *J. Am. Med. Assoc.* **323**, 1439–1440 (2020).
21. Halpern, N. A., Goldman, D. A., Tan, K. S. & Pastores, S. M. Trends in critical care beds and use among population groups and Medicare and Medicaid beneficiaries in the United States: 2000–2010. *Crit. Care Med.* **44**, 1490–1499 (2016).
22. Halpern, N. A. & Pastores, S. M. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit. Care Med.* **38**, 65–71 (2010).
23. Hermans, G. et al. Acute outcomes and 1-year mortality of intensive care unit-acquired weakness. A cohort study and propensity-matched analysis. *Am. J. Respir. Crit. Care Med.* **190**, 410–420 (2014).
24. Zhang, L. et al. Early mobilization of critically ill patients in the intensive care unit: a systematic review and meta-analysis. *PLoS ONE* **14**, e0223185 (2019).
25. Donchin, Y. et al. A look into the nature and causes of human errors in the intensive care unit. *Crit. Care Med.* **23**, 294–300 (1995).
26. Hodgson, C. L., Berney, S., Harrold, M., Saxena, M. & Bellomo, R. Clinical review: early patient mobilization in the ICU. *Crit. Care* **17**, 207 (2013).
27. Verceles, A. C. & Hager, E. R. Use of accelerometry to monitor physical activity in critically ill subjects: a systematic review. *Respir. Care* **60**, 1330–1336 (2015).
28. Ma, A. J. et al. Measuring patient mobility in the ICU using a novel noninvasive sensor. *Crit. Care Med.* **45**, 630–636 (2017).
29. Yeung, S. et al. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *NPJ Digit. Med.* **2**, 11 (2019).
This study used computer vision to simultaneously categorize patient mobilization activities in intensive care units and count the number of healthcare personnel involved in each activity.
30. Davoudi, A. et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Sci. Rep.* **9**, 8020 (2019).
This study used cameras and wearable sensors to track the physical movement of delirious and non-delirious patients in an intensive care unit.
31. WHO. *Report on the Burden of Endemic Health Care-associated Infection Worldwide* https://apps.who.int/iris/handle/10665/80135 (2011).
32. Vincent, J.-L. Nosocomial infections in adult intensive-care units. *Lancet* **361**, 2068–2077 (2003).
33. Gould, D. J., Moralejo, D., Drey, N., Chudleigh, J. H. & Taljaard, M. Interventions to improve hand hygiene compliance in patient care. *Cochrane Database Syst. Rev.* **9**, CD005186 (2017).
34. Srigley, J. A., Furness, C. D., Baker, G. R. & Gardam, M. Quantification of the Hawthorne effect in hand hygiene compliance monitoring using an electronic monitoring system: a retrospective cohort study. *BMJ Qual. Saf.* **23**, 974–980 (2014).
35. Shirehjini, A. A. N., Yassine, A. & Shirmohammadi, S. Equipment location in hospitals using RFID-based positioning system. *IEEE Trans. Inf. Technol. Biomed.* **16**, 1058–1069 (2012).
36. Sax, H. et al. 'My five moments for hand hygiene': a user-centred design approach to understand, train, monitor and report hand hygiene. *J. Hosp. Infect.* **67**, 9–21 (2007).
37. Haque, A. et al. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. In *Proc. 2nd Machine Learning for Healthcare Conference* 75–87 (PMLR, 2017).
This study evaluated the performance of depth sensors and covert auditors at measuring hand hygiene compliance in a hospital unit.
38. Singh, A. et al. Automatic detection of hand hygiene using computer vision technology. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1093/jamia/ocaa115 (2020).
39. Chen, J., Cremer, J. F., Zarei, K., Segre, A. M. & Polgreen, P. M. Using computer vision and depth sensing to measure healthcare worker-patient contacts and personal protective equipment adherence within hospital rooms. *Open Forum Infect. Dis.* **3**, ofv200 (2016).
40. Awwad, S., Tarvade, S., Piccardi, M. & Gattas, D. J. The use of privacy-protected computer vision to measure the quality of healthcare worker hand hygiene. *Int. J. Qual. Health Care* **31**, 36–42 (2019).
41. Weiser, T. G. et al. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet* **372**, 139–144 (2008).
42. Anderson, O., Davis, R., Hanna, G. B. & Vincent, C. A. Surgical adverse events: a systematic review. *Am. J. Surg.* **206**, 253–262 (2013).
43. Bonrath, E. M., Dedy, N. J., Gordon, L. E. & Grantcharov, T. P. Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. *Ann. Surg.* **262**, 205–212 (2015).
44. Vaidya, A. et al. Current status of technical skills assessment tools in surgery: a systematic review. *J. Surg. Res.* **246**, 342–378 (2020).
45. Ghasemloonia, A. et al. Surgical skill assessment using motion quality and smoothness. *J. Surg. Educ.* **74**, 295–305 (2017).
46. Khalid, S., Goldenberg, M., Grantcharov, T., Taati, B. & Rudzicz, F. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA Netw. Open* **3**, e201664 (2020).
47. Law, H., Ghani, K. & Deng, J. Surgeon Technical skill assessment using computer vision based analysis. In *Proc. 2nd Machine Learning for Healthcare Conference* 88–99 (PMLR, 2017).
48. Jin, A. et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *Proc. Winter Conference on Applications of Computer Vision* 691–699 (IEEE, 2018).
49. Twinanda, A. P. et al. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **36**, 86–97 (2017).
50. Hashimoto, D. A., Rosman, G., Rus, D. & Meireles, O. R. Artificial intelligence in surgery: promises and perils. *Ann. Surg.* **268**, 70–76 (2018).
51. Greenberg, C. C., Regenbogen, S. E., Lipsitz, S. R., Diaz-Flores, R. & Gawande, A. A. The frequency and significance of discrepancies in the surgical count. *Ann. Surg.* **248**, 337–341 (2008).
52. Agrawal, A. Counting matters: lessons from the root cause analysis of a retained surgical item. *Jt. Comm. J. Qual. Patient Saf.* **38**, 566–574 (2012).
53. Hempel, S. et al. Wrong-site surgery, retained surgical items, and surgical fires: a systematic review of surgical never events. *JAMA Surg.* **150**, 796–805 (2015).
54. Cima, R. R. et al. Using a data-matrix-coded sponge counting system across a surgical practice: impact after 18 months. *Jt. Comm. J. Qual. Patient Saf.* **37**, 51–58 (2011).
55. Rupp, C. C. et al. Effectiveness of a radiofrequency detection system as an adjunct to manual counting protocols for tracking surgical sponges: a prospective trial of 2,285 patients. *J. Am. Coll. Surg.* **215**, 524–533 (2012).
56. Kassahun, Y. et al. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int. J. Comput. Assist. Radiol. Surg.* **11**, 553–568 (2016).
57. Kadkhodamohammadi, A., Gangi, A., de Mathelin, M. & Padoy, N. A multi-view RGB-D approach for human pose estimation in operating rooms. In *Proc. Winter Conference on Applications of Computer Vision* 363–372 (IEEE, 2017).
58. Jung, J. J., Jüni, P., Lebovic, G. & Grantcharov, T. First-year analysis of the operating room black box study. *Ann. Surg.* **271**, 122–127 (2020).
59. Joukes, E., Abu-Hanna, A., Cornet, R. & de Keizer, N. F. Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record. *Appl. Clin. Inform.* **9**, 46–53 (2018).
60. Heaton, H. A., Castaneda-Guarderas, A., Trotter, E. R., Erwin, P. J. & Bellolio, M. F. Effect of scribes on patient throughput, revenue, and patient and provider satisfaction: a systematic review and meta-analysis. *Am. J. Emerg. Med.* **34**, 2018–2028 (2016).
61. Rich, N. The impact of working as a medical scribe. *Am. J. Emerg. Med.* **35**, 513 (2017).
62. Boulton, C. How Google Glass automates patient documentation for dignity health. *Wall Street Journal* (16 June 2014).

63. Blackley, S. V., Huynh, J., Wang, L., Korach, Z. & Zhou, L. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *J. Am. Med. Inform. Assoc.* **26**, 324–338 (2019).
64. Chiu, C.-C. et al. Speech recognition for medical conversations. In *Proc. 18th Annual Conference of the International Speech Communication Association* 2972–2976 (ISCA, 2018).
This paper developed a speech-recognition algorithm to transcribe anonymized conversations between patients and clinicians.
65. Pranaat, R. et al. Use of simulation based on an electronic health records environment to evaluate the structure and accuracy of notes generated by medical scribes: proof-of-concept study. *JMIR Med. Inform.* **5**, e30 (2017).
66. Kaplan, R. S. et al. Using time-driven activity-based costing to identify value improvement opportunities in healthcare. *J. Healthc. Manag.* **59**, 399–412 (2014).
67. Porter, M. E. Value-based health care delivery. *Ann. Surg.* **248**, 503–509 (2008).
68. Keel, G., Savage, C., Rafiq, M. & Mazzocato, P. Time-driven activity-based costing in health care: a systematic review of the literature. *Health Policy* **121**, 755–763 (2017).
69. French, K. E. et al. Measuring the value of process improvement initiatives in a preoperative assessment center using time-driven activity-based costing. *Healthcare* **1**, 136–142 (2013).
70. Sánchez, D., Tentori, M. & Favela, J. Activity recognition for the smart hospital. *IEEE Intelligent Systems* **23**, 50–57 (2008).
71. United Nations. *World Population Ageing 2019* https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2019_worldpopulationageing_report.pdf (2020).
72. Mamikonian-Zarpas, A. & Laganá, L. The relationship between older adults' risk for a future fall and difficulty performing activities of daily living. *J. Aging Gerontol.* **3**, 8–16 (2015).
73. Stineman, M. G. et al. All-cause 1-, 5-, and 10-year mortality in elderly people according to activities of daily living stage. *J. Am. Geriatr. Soc.* **60**, 485–492 (2012).
74. Phelan, E. A., Williams, B., Penninx, B. W. J. H., LoGerfo, J. P. & Leveille, S. G. Activities of daily living function and disability in older adults in a randomized trial of the health enhancement program. *J. Gerontol. A* **59**, M838–M843 (2004).
75. Carlsson, G., Haak, M., Nygren, C. & Iwarsson, S. Self-reported versus professionally assessed functional limitations in community-dwelling very old individuals. *Int. J. Rehabil. Res.* **35**, 299–304 (2012).
76. Wang, Z., Yang, Z. & Dong, T. A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time. *Sensors* **17**, 341 (2017).
77. Katz, S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *J. Am. Geriatr. Soc.* **31**, 721–727 (1983).
78. Uddin, M. Z., Khaksar, W. & Torresen, J. Ambient sensors for elderly care and independent living: a survey. *Sensors* **18**, 2027 (2018).
79. Luo, Z. et al. Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. In *Proc. 3rd Machine Learning for Healthcare Conference* 1–18 (PMLR, 2018).
This study created spatial and temporal summaries of activities of daily living using a depth and thermal sensor inside the bedroom of an older resident.
80. Cheng, H., Liu, Z., Zhao, Y., Ye, G. & Sun, X. Real world activity summary for senior home monitoring. *Multimedia Tools Appl.* **70**, 177–197 (2014).
81. Lee, M.-T., Jang, Y. & Chang, W.-Y. How do impairments in cognitive functions affect activities of daily living functions in older adults? *PLoS ONE* **14**, e0218112 (2019).
82. Chen, J., Zhang, J., Kam, A. H. & Shue, L. An automatic acoustic bathroom monitoring system. In *Proc. International Symposium on Circuits and Systems* 1750–1753 (IEEE, 2005).
83. Shrestha, A. et al. Elderly care: activities of daily living classification with an S band radar. *J. Eng.* **2019**, 7601–7606 (2019).
84. Ganz, D. A. & Latham, N. K. Prevention of falls in community-dwelling older adults. *N. Engl. J. Med.* **382**, 734–743 (2020).
85. Bergen, G., Stevens, M. R. & Burns, E. R. Falls and fall injuries among adults aged ≥65 years — United States, 2014. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 993–998 (2016).
86. Wild, D., Nayak, U. S. & Isaacs, B. How dangerous are falls in old people at home? *Br. Med. J. (Clin. Res. Ed.)* **282**, 266–268 (1981).
87. Scheffer, A. C., Schuurmans, M. J., van Dijk, N., van der Hooft, T. & de Rooij, S. E. Fear of falling: measurement strategy, prevalence, risk factors and consequences among older persons. *Age Ageing* **37**, 19–24 (2008).
88. Pol, M. et al. Older people's perspectives regarding the use of sensor monitoring in their home. *Gerontologist* **56**, 485–493 (2016).
89. Erol, B., Amin, M. G. & Boashash, B. Range-Doppler radar sensor fusion for fall detection. In *Proc. IEEE Radar Conference* 819–824 (IEEE, 2017).
90. Chaudhuri, S., Thompson, H. & Demiris, G. Fall detection devices and their use with older adults: a systematic review. *J. Geriatr. Phys. Ther.* **37**, 178–196 (2014).
91. Tegou, T. et al. A low-cost indoor activity monitoring system for detecting frailty in older adults. *Sensors* **19**, 452 (2019).
92. Rantz, M. et al. Automated in-home fall risk assessment and detection sensor system for elders. *Gerontologist* **55**, S78–S87 (2015).
93. Su, B. Y., Ho, K. C., Rantz, M. J. & Skubic, M. Doppler radar fall activity detection using the wavelet transform. *IEEE Trans. Biomed. Eng.* **62**, 865–875 (2015).
94. Stone, E. E. & Skubic, M. Fall detection in homes of older adults using the Microsoft Kinect. *IEEE J. Biomed. Health Inform.* **19**, 290–301 (2015).
95. Rantz, M. et al. Randomized trial of intelligent sensor system for early illness alerts in senior housing. *J. Am. Med. Dir. Assoc.* **18**, 860–870 (2017).
This randomized trial investigated the clinical efficacy of a real-time intervention system—triggered by abnormal gait patterns, as detected by ambient sensors—on the walking ability of older individuals at home.
96. Kwolek, B. & Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput. Methods Programs Biomed.* **117**, 489–501 (2014).
97. Wren, T. A. L., Gorton, G. E. III, Ounpuu, S. & Tucker, C. A. Efficacy of clinical gait analysis: a systematic review. *Gait Posture* **34**, 149–153 (2011).
98. Wren, T. A. et al. Outcomes of lower extremity orthopedic surgery in ambulatory children with cerebral palsy with and without gait analysis: results of a randomized controlled trial. *Gait Posture* **38**, 236–241 (2013).
99. Del Din, S. et al. Gait analysis with wearables predicts conversion to Parkinson disease. *Ann. Neurol.* **86**, 357–367 (2019).
100. Kidziński, Ł., Delp, S. & Schwartz, M. Automatic real-time gait event detection in children using deep neural networks. *PLoS ONE* **14**, e0211466 (2019).
101. Díaz, S., Stephenson, J. B. & Labrador, M. A. Use of wearable sensor technology in gait, balance, and range of motion analysis. *Appl. Sci.* **10**, 234 (2020).
102. Juen, J., Cheng, Q., Prieto-Centurion, V., Krishnan, J. A. & Schatz, B. Health monitors for chronic disease by gait analysis with mobile phones. *Telemed. J. E Health* **20**, 1035–1041 (2014).
103. Kononova, A. et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. *JMIR Mhealth Uhealth* **7**, e9832 (2019).
104. Da Gama, A., Fallavollita, P., Teichrieb, V., & Navab, N. Motor rehabilitation using Kinect: a systematic review. *Games Health J.* **4**, 123–135 (2015).
105. Cho, C.-W., Chao, W.-H., Lin, S.-H. & Chen, Y.-Y. A vision-based analysis system for gait recognition in patients with Parkinson's disease. *Expert Syst. Appl.* **36**, 7033–7039 (2009).
106. Seifert, A., Zoubir, A. M. & Amin, M. G. Detection of gait asymmetry using indoor Doppler radar. In *Proc. IEEE Radar Conference* 1–6 (IEEE, 2019).
107. Altaf, M. U. B., Butko, T., Juang, B. H. & Juang, B.-H. Acoustic gaits: gait analysis with footstep sounds. *IEEE Trans. Biomed. Eng.* **62**, 2001–2011 (2015).
108. Galna, B. et al. Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait Posture* **39**, 1062–1068 (2014).
109. Jaume-i-Capó, A., Martínez-Bueso, P., Moyà-Alcover, B. & Varona, J. Interactive rehabilitation system for improvement of balance therapies in people with cerebral palsy. *IEEE Trans. Neural Syst. Rehabil. Eng.* **22**, 419–427 (2014).
110. Tinetti, M. E., Williams, T. F. & Mayewski, R. Fall risk index for elderly patients based on number of chronic disabilities. *Am. J. Med.* **80**, 429–434 (1986).
111. Wang, C. et al. Multimodal gait analysis based on wearable inertial and microphone sensors. In *Proc. IEEE SmartWorld* 1–8 (2017).
112. Mental Health America. *Mental Health in America - Adult Data 2018* https://www.mhanational.org/issues/mental-health-america-adult-data-2018 (2018).
113. Wittchen, H. U. et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur. Neuropsychopharmacol.* **21**, 655–679 (2011).
114. Snowden, L. R. Bias in mental health assessment and intervention: theory and evidence. *Am. J. Public Health* **93**, 239–243 (2003).
115. Shatte, A. B. R., Hutchinson, D. M. & Teague, S. J. Machine learning in mental health: a scoping review of methods and applications. *Psychol. Med.* **49**, 1426–1448 (2019).
116. Chakraborty, D. et al. Assessment and prediction of negative symptoms of schizophrenia from RGB+ D movement signals. In *Proc. 19th International Workshop on Multimedia Signal Processing* 1–6 (2017).
117. Pestian, J. P. et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life Threat. Behav.* **46**, 154–159 (2016).
118. Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S. & Stiles, W. B. Therapist effects in outpatient psychotherapy: a three-level growth curve approach. *J. Couns. Psychol.* **54**, 32–39 (2007).
119. Miner, A. S. et al. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digit. Med.* **3**, 82 (2020).
120. Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C. & Narayanan, S. S. "Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE* **10**, e0143055 (2015).
121. Ewbank, M. P. et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry* **77**, 35–43 (2020).
122. Sadeghian, A., Alahi, A. & Savarese, S. Tracking the untrackable: learning to track multiple cues with long-term dependencies. In *Proc. Conference on Computer Vision and Pattern Recognition* 300–311 (IEEE, 2017).
123. Liu, G. et al. Image inpainting for irregular holes using partial convolutions. In *Proc. 15th European Conference on Computer Vision* 89–105 (Springer, 2018).
124. Marafioti, A., Perraudin, N., Holighaus, N. & Majdak, P. A context encoder for audio inpainting. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 2362–2372 (2019).
125. Chen, Y., Tian, Y. & He, M. Monocular human pose estimation: a survey of deep learning-based methods. *Comput. Vis. Image Underst.* **192**, 102897 (2020).
126. Krishna, R. et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **123**, 32–73 (2017).
127. Johnson, J. et al. Image retrieval using scene graphs. In *Proc. Conference on Computer Vision and Pattern Recognition* 3668–3678 (IEEE, 2015).
128. Shi, J., Zhang, H. & Li, J. Explainable and explicit visual reasoning over scene graphs. In *Proc. Conference on Computer Vision and Pattern Recognition* 8368–8376 (IEEE/CVF, 2019).
129. Halamka, J. D. Early experiences with big data at an academic medical center. *Health Aff.* **33**, 1132–1138 (2014).
130. Verbraeken, J. et al. A survey on distributed machine learning. *ACM Comput. Surv.* **53**, 30 (2020).
131. You, Y. et al. Large batch optimization for deep learning: training BERT in 76 minutes. In *Proc. 8th International Conference on Learning Representations* 1–38 (2020).
132. Kitaev, N., Kaiser, Ł. & Levskaya, A. Reformer: the efficient transformer. In *Proc. 8th International Conference on Learning Representations* 1–12 (2020).
133. Heilbron, F., Niebles, J. & Ghanem, B. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proc. Conference on Computer Vision and Pattern Recognition* 1914–1923 (IEEE, 2016).

134. Zhu, Y., Lan, Z., Newsam, S. & Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In *Proc. 14th Asian Conference on Computer Vision* 363–378 (Springer, 2019).
135. Han, S., Mao, H. & Dally, W. J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In *Proc. 4th International Conference on Learning Representations* 1–14 (2016).
This paper introduced a method to compress neural network models and reduce their computational and storage requirements.
136. Micikeviciusd, P. et al. Mixed precision training. In *Proc. 6th International Conference on Learning Representations* 1–12 (2018).
137. Yu, G. & Yuan, J. Fast action proposals for human action detection and search. In *Proc. Conference on Computer Vision and Pattern Recognition* 1302–1311 (IEEE, 2015).
138. Zou, J. & Schiebinger, L. AI can be sexist and racist — it's time to make it fair. *Nature* **559**, 324–326 (2018).
139. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. *Adv. Neural Inf. Process. Syst*. **30**, 5680–5689 (2017).
140. Neyshabur, B., Bhojanapalli, S., McAllester, D. & Srebro, N. Exploring generalization in deep learning. *Adv. Neural Inf. Process. Syst*. **30**, 5947–5956 (2017).
141. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* 328–339 (2018).
142. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. **22**, 1345–1359 (2010).
143. Patel, V. M., Gopalan, R., Li, R. & Chellappa, R. Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag*. **32**, 53–69 (2015).
144. Wang, Y., Kwok, J., Ni, L. M. & Yao, Q. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv*. **53**, 63 (2020).
145. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell*. **1**, 389–399 (2019).
146. Li, C., Lubecke, V. M., Boric-Lubecke, O. & Lin, J. A review on recent advances in Doppler radar sensors for noncontact healthcare monitoring. *IEEE Trans. Microw. Theory Tech*. **61**, 2046–2060 (2013).
147. Rockhold, F., Nisen, P. & Freeman, A. Data sharing at a crossroads. *N. Engl. J. Med*. **375**, 1115–1117 (2016).
148. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med*. **25**, 1337–1340 (2019).
149. El Emam, K., Jonker, E., Arbuckle, L. & Malin, B. A systematic review of re-identification attacks on health data. *PLoS ONE* **6**, e28071 (2011).
150. Nasrollahi, K. & Moeslund, T. Super-resolution: a comprehensive survey. *Mach. Vis. Appl*. **25**, 1423–1468 (2014).
151. Brewster, T. How an amateur rap crew stole surveillance tech that tracks almost every American. *Forbes Magazine* (12 October 2018).
152. Cutler, J. E. How can patients make money off their medical data? *Bloomberg Law* (29 January 2019).
153. Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S. & Rubin, D. L. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit. Med*. **2**, 78 (2019).
154. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med*. **169**, 866–872 (2018).
155. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care — addressing ethical challenges. *N. Engl. J. Med*. **378**, 981–983 (2018).
156. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* 77–91 (2018).
157. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* **21**, E167–E179 (2019).
158. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med*. **170**, 51–58 (2019).
159. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
This article proposed a framework for evaluating model interpretability through predictive accuracy, descriptive accuracy and relevancy.
160. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med*. **25**, 30–36 (2019).
161. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med*. **162**, 55–63 (2015).
162. Mitchell, M. et al. Model cards for model reporting. In *Proc. 2nd Conference on Fairness, Accountability, and Transparency* 220–229 (2019).
163. Thomas, R. et al. Deliberative democracy and cancer screening consent: a randomised control trial of the effect of a community jury on men's knowledge about and intentions to participate in PSA screening. *BMJ Open* **4**, e005691 (2014).
164. Otto, J. L., Holodniy, M. & DeFraites, R. F. Public health practice is not research. *Am. J. Public Health* **104**, 596–602 (2014).
165. Gerke, S., Yeung, S. & Cohen, I. G. Ethical and legal aspects of ambient intelligence in hospitals. *J. Am. Med. Assoc*. **323**, 601–602 (2020).
166. Kim, J. W., Jang, B. & Yoo, H. Privacy-preserving aggregation of personal health data streams. *PLoS ONE* **13**, e0207639 (2018).
167. van der Maaten, L., Postma, E. & van den Herik, J. Dimensionality reduction: a comparative. *J. Mach. Learn. Res*. **10**, 13 (2009).
168. Kocabas, M., Athanasiou, N. & Black, M. J. VIBE: video inference for human body pose and shape estimation. In *Proc. Conference on Computer Vision and Pattern Recognition* 5253–5263 (IEEE/CVF, 2020).
169. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
This paper proposed federated learning, a method for training a shared model while the data is distributed across multiple client devices.
170. Gentry, C. Fully homomorphic encryption using ideal lattices. In *Proc 41st Symposium on Theory of Computing* 169–178 (ACM, 2009).
This paper proposed the first fully homomorphic encryption scheme that supports addition and multiplication on encrypted data.
171. McCoy, S. T. *Aboard USNS Comfort* (US Navy, 2003).

Acknowledgements We thank A. Kaushal, D. C. Magnus, G. Burke, K. Schulman and M. Hutson for providing comments on this paper. We also thank our clinical collaborators over the years, including A. S. Miner, A. Singh, B. Campbell, D. F. Amanatullah, F. R. Salipur, H. Rubin, J. Jopling, K. Deru, N. L. Downing, R. Nazerali, T. Platchek and W. Beninati, and our technical collaborators over the years, including A. Alahi, A. Rege, B. Liu, B. Peng, D. Zhao, E. Chou, E. Adeli, G. M. Bianconi, G. Pusiol, H. Cai, J. Beal, J.-T. Hsieh, M. Guo, R. Mehra, S. Mehra, S. Yeung and Z. Luo. A.H.'s graduate work was partially supported by the US Office of Naval Research (grant N00014-16-1-2127) and the Stanford Institute for Human-Centered Artificial Intelligence.

Author contributions A.H., A.M. and L.F-F. conceptualized the paper and its structure. A.H. and L.F.-F. wrote the paper. A.H. created the figures. A.M. provided substantial additions and edits. All authors contributed to multiple parts of the paper, as well as the final style and overall content.

Competing interests A.M. has financial interests in Prealize Health. L.F.-F. and A.M. have financial interests in Dawnlight Technologies. A.H. declares no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2669-y.
Correspondence and requests for materials should be addressed to L.F.-F.
Peer review information *Nature* thanks Andrew Beam, Eric Topol and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article

The Magellanic Corona as the key to the formation of the Magellanic Stream

https://doi.org/10.1038/s41586-020-2663-4

Received: 16 January 2020

Accepted: 5 June 2020

Published online: 9 September 2020

Check for updates

S. Lucchini[1], E. D'Onghia[1,2,3]✉, A. J. Fox[4], C. Bustard[1], J. Bland-Hawthorn[5,6] & E. Zweibel[1,2]

The dominant gaseous structure in the Galactic halo is the Magellanic Stream. This extended network of neutral and ionized filaments surrounds the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC), the two most massive satellite galaxies of the Milky Way[1–4]. Recent observations indicate that the LMC and SMC are on their first passage around the Galaxy[5], that the Magellanic Stream is made up of gas stripped from both clouds[2,6,7] and that the majority of this gas is ionized[8,9]. Although it has long been suspected that tidal forces[10,11] and ram-pressure stripping[12,13] contributed to the formation of the Magellanic Stream, models have not been able to provide a full understanding of its origins[3]. Several recent developments—including the discovery of dwarf galaxies associated with the Magellanic group[14–16], determination of the high mass of the LMC[17], detection of highly ionized gas near stars in the LMC[18,19] and predictions of cosmological simulations[20,21]—support the existence of a halo of warm (roughly 500,000 kelvin) ionized gas around the LMC (the 'Magellanic Corona'). Here we report that, by including this Magellanic Corona in hydrodynamic simulations of the Magellanic Clouds falling onto the Milky Way, we can reproduce the Magellanic Stream and its leading arm. Our simulations explain the filamentary structure, spatial extent, radial-velocity gradient and total ionized-gas mass of the Magellanic Stream. We predict that the Magellanic Corona will be unambiguously observable via high-ionization absorption lines in the ultraviolet spectra of background quasars lying near the LMC.

The most successful model so far of the formation of the Magellanic Stream is known as the first-infall model[10,22]. In this model, tidal forces from the LMC acting on the SMC when these clouds are at their first pericentric passage around the Milky Way lead to the formation of the stream. This model is motivated by the high tangential velocities of the clouds[5] and the strong morphological disturbances observed in the SMC[10,11], and it successfully reproduces the size and shape of the Magellanic Stream. However, several issues remain[3]: (i) the observed stream is much more extended spatially and up to ten times more massive than the simulated stream, especially when including its ionized component, which dominates the mass budget[8]; (ii) the fragmented structure of the stream and leading arm[4,23] indicates that interaction with the Milky Way's gas corona is important and should be included; and (iii) the stream is bifurcated, with kinematic and chemical analyses indicating that gas from both the LMC and SMC is present[2,6,7]. This indicates that the Magellanic Stream has a dual origin, whereas tidal models predict an SMC origin because of the shallower potential well of the SMC.

By including the Magellanic Corona in hydrodynamic simulations of stream formation, we resolve the mass-budget discrepancy of the stream and, crucially, reproduce the ionized component. The Magellanic Corona appears to be the key missing ingredient in models of stream formation. Our simulations were run using the GIZMO hydrodynamic *N*-body code. They incorporated radiative cooling and heating, star formation and stellar feedback to model the LMC–SMC–Milky Way dynamics, including the Magellanic Corona (see Methods). During the initial stages of the LMC–SMC tidal interaction, the pair lie outside the gravitational influence of the Milky Way. The cold gas in the extended disk of the SMC is tidally stripped through repeated encounters with the LMC (as illustrated in Extended Data Figs. 2, 3b and Supplementary Video 3) that occur over a period of 5.7 Gyr. Because the model includes more massive and more extended disks for the clouds than did previous studies[10], these repeated orbits of the SMC around the LMC also result in gas extraction from the LMC[11] by dwarf–dwarf galaxy interaction. However, this process acting on both clouds contributes only 10%–20% of the total stream mass.

During the early period, before the LMC–SMC pair fell into the Milky Way, a Magellanic Corona of gas with temperature $T \approx 5 \times 10^5$ K and mass $M \approx 3 \times 10^9 M_\odot$ (where $M_\odot$ is the mass of the Sun) surrounded the Magellanic system and extended out to the virial radius of the LMC (100 kpc). This corona removes cold gas from the outer disk of the SMC and heats it up by compression, as illustrated in Extended Data Fig. 3d. Later, the corona provides an additional source of ionized gas that contributes to the total mass in the Magellanic Stream. The Magellanic Corona is therefore a source of pressure, heating and mass.

Once the clouds fell into the Milky Way and its hot corona, the stream was amplified by the Milky Way potential until it extended over 200

[1]Department of Physics, University of Wisconsin-Madison, Madison, WI, USA. [2]Department of Astronomy, University of Wisconsin-Madison, Madison, WI, USA. [3]Center for Computational Astrophysics, Flatiron Institute, New York, NY, USA. [4]AURA for ESA, Space Telescope Science Institute, Baltimore, MD, USA. [5]Sydney Institute for Astronomy, School of Physics, University of Sydney, Sydney, New South Wales, Australia. [6]ARC Centre of Excellence for All Sky Astrophysics in 3D, Canberra, Australian Capital Territory, Australia. ✉e-mail: edonghia@astro.wisc.edu

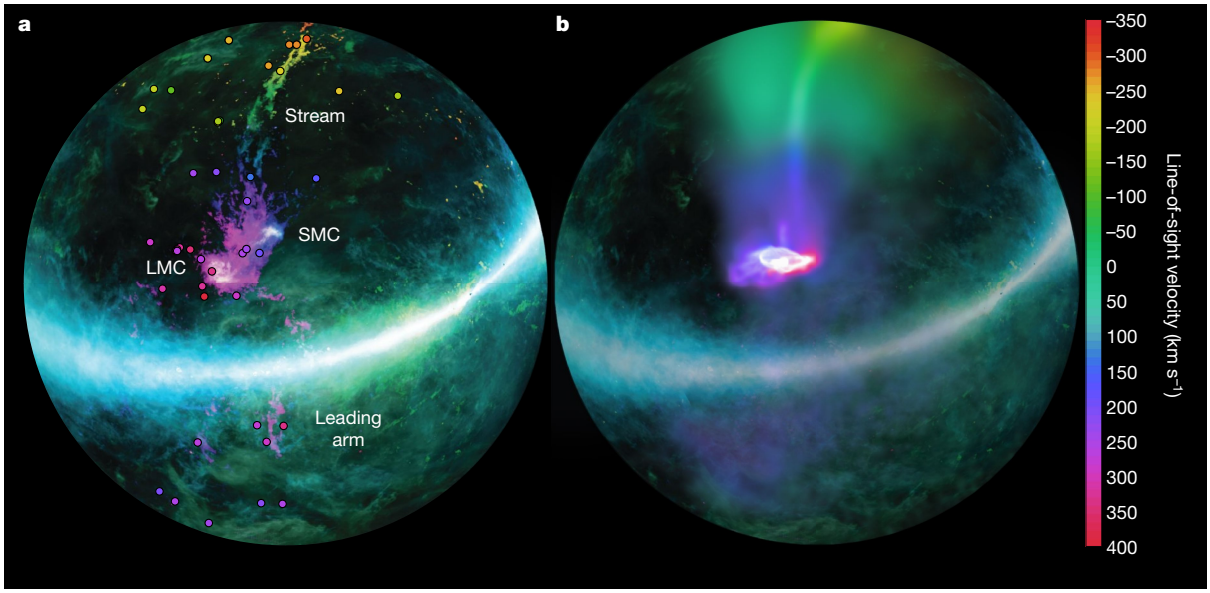


Fig. 1 | The Magellanic Stream in zenithal equal-area coordinates. a, Observed H I data[24] for the Magellanic Stream, with the line-of-sight velocity indicated by the colour scale (from −350 km s^{-1} to 400 km s^{-1}) and the relative gas column density indicated by the brightness. The points represent the sightlines with ultraviolet-absorption-line observations from the Hubble Space Telescope[8], coloured by their line-of-sight velocity. These points show the extent of the ionized gas associated with the stream. **b**, The results of the model including the Magellanic Corona and the Milky Way's hot corona. Gas originating in both the LMC and the SMC disks is shown in the model, without separating neutral gas from ionized gas. This affects the morphology of the stream, causing the model to appear smoother and less fragmented than the data. However, the model reproduces the current spatial location and velocity of both clouds, and the velocity gradient of the gas along the stream. The Milky Way disk and background are extracted from real H I images[24]. Image in **a** adapted with permission from ref. [24], American Astronomical Society.

degrees in the sky, with both leading and trailing components. Figure 1 shows the Magellanic Stream at the present time in zenithal equal-area projection in the numerical experiment (Fig. 1b), compared with the observed stream (Fig. 1a)[23,24]. Figure 2 displays the simulated stream in Magellanic coordinates, showing the neutral disk gas stripped off the clouds (Fig. 2b), the neutral gas and the ionized Magellanic Corona gas (Fig. 2a), and the line-of-sight velocity along the stream (Fig. 2c). These are shown in comparison to observations, which are represented by greyscale contours in Fig. 2b, c[23]. (See Supplementary Videos 1 and 2 for videos of the infalling clouds.) The Milky Way's hot corona included in this model has a total mass of around $2 \times 10^9 M_\odot$ and does not rotate (see Methods). The presence of the hot Milky Way gas and the Magellanic Corona have a large effect on the kinematics of the stream. To illustrate this, Fig. 1b displays a comparison of line-of-sight velocities of the stream with the H I velocity gradient observed[23] in the case when both the Magellanic and Milky Way coronae are included. The model shows a kinematic gradient from negative to positive velocities along the stream (from −350 km s^{-1} to 400 km s^{-1}), in good agreement with the observed data (Figs. 1a, 2c). Whereas previous models found the gas to be moving roughly 100 km s^{-1} faster than observations in the leading arm[11] and slower in the stream[10], the inclusion of coronal gas decelerates the leading arm to better match the observed velocity gradient. The remaining offset in velocity between the observations and the model at the tail of the stream may be resolved by modifying the orbits of the clouds around the Milky Way. However, the cold gas column density in this region is smoother than in observations, which indicates that the leading arm is clumpy and fragmented[4,23] (see Fig. 2b).

In our model, both the LMC and the SMC contribute to the formation of the Magellanic Stream. Most of the gas is pulled from the SMC, but there is also a tenuous filamentary contribution from the LMC, produced by tidal interactions with the Milky Way and ram-pressure stripping in its hot corona. When the Magellanic system first falls into the Milky Way, the Magellanic Corona is extended. Under the influence of the gravitational potential of the Milky Way, about 22% of the initial mass of the Magellanic Corona becomes unbound from the LMC and incorporated into the stream. Thus, by mixing with the underlying hot gas of the Milky Way, the Magellanic Corona contributes to the large ionized mass of the stream. Figure 3 shows that the Magellanic Corona contributes around 50% of the mass in the leading arm and more than 50% of the total ionized mass in the stream. The other roughly 50% of the mass (in both the leading arm and the stream) is composed of gas extracted earlier from the SMC by its mutual interaction with the LMC, with some gas heated by the Magellanic Corona before infall. This additional source of ionized gas has not been accounted for in previous theoretical work, and reconciles the mass budget for the Magellanic Stream.

Another outcome of the model concerns the survivability of the Magellanic Stream and its leading arm in the presence of the Milky Way's hot corona. H I studies[4,25] show that the leading arm is fragmented, as expected from simulations of its passage through the Galactic halo[26], yet it survives. However, recent hydrodynamic simulations have challenged the overall survivability of the leading arm when the Milky Way's hot corona is included[27]. The numerical experiment reported here shows that the leading arm survives if the hot Milky Way halo has a density of $n \approx 1.7 \times 10^{-5}$ cm^{-3} at a distance of 50 kpc from Galactic centre (see Extended Data Fig. 1). While the Milky Way's corona regulates the formation and morphology of the leading arm, the inclusion of the Magellanic Corona affects its spatial extent (see Extended Data Fig. 4). The warm gas surrounding the clouds provides a shield around the stripped gas to allow the leading-arm gas to penetrate further into the Milky Way's hot corona. Even if the leading arm turns out to have a non-Magellanic origin, as suggested recently[27], the inclusion of the Magellanic Corona still provides the bulk of the mass of the trailing stream, including its ionized component.

The inclusion of the Magellanic Corona is further supported by a recent estimate of the ambient gas density near the leading arm[28]. Following the discovery of stars formed in situ in the leading arm[29], a recent study[28] reports that the density of coronal gas required to separate these young stars from their proposed gaseous nursery (the region known as leading arm II) is an order of magnitude higher than existing

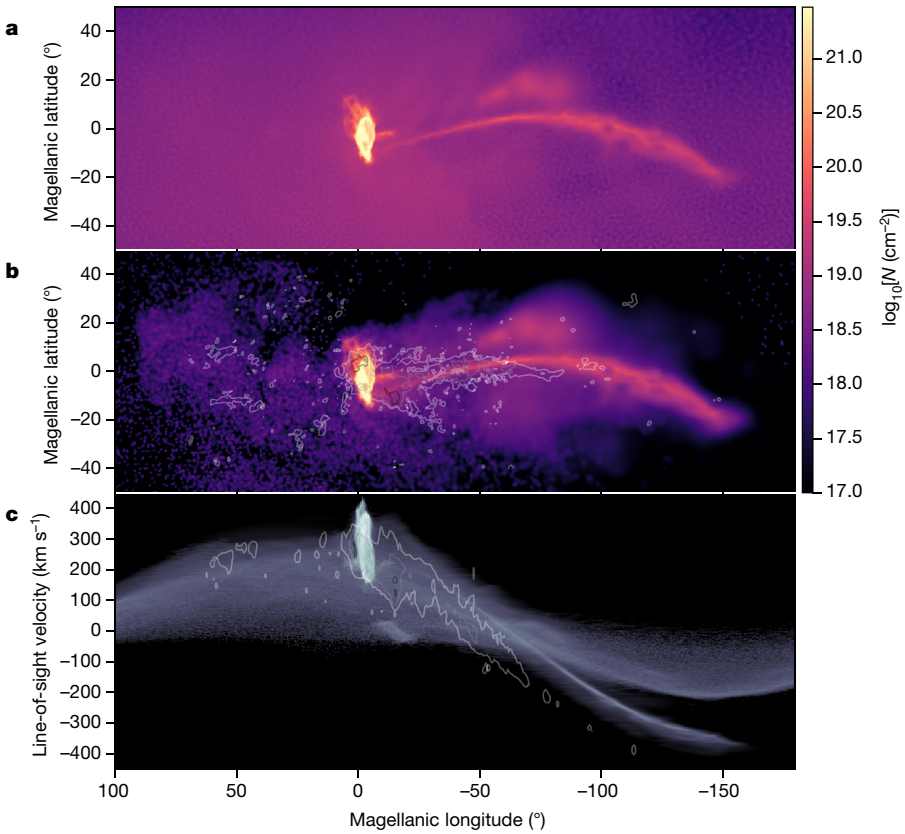


Fig. 2 | Gas column density and velocity in Magellanic coordinates. a, The gas column density *N* of the simulated stream, which is composed of the Magellanic Corona gas and cold disk gas stripped from the clouds, displayed in Magellanic coordinates. **b**, Column density of only the simulated cold gas stream, compared to H I data[23]. Black, grey and white contours corresponding to observed column densities of 10^{19} cm^{-2}, 10^{20} cm^{-2} and 10^{21} cm^{-2}, respectively. **c**, The line-of-sight velocity of the total stream gas as a function of Magellanic longitude, with contours as in **b** and brightness showing the relative density.

measurements of the coronal density of the Milky Way[30]. This discrepancy can be resolved by taking the Magellanic Corona into account, because the Magellanic Corona can add to the Milky Way's corona to yield the high total density needed to ram-pressure-strip the leading arm II region away from the nascent stars.

An additional consequence of this model is a possible explanation for the lack of a stellar component of the stream. In tidal models, stars (in addition to gas) should be stripped from both clouds as a result of the gravitational interactions they experience before falling into the Milky Way. Such a stellar stream has yet to be discovered, even though

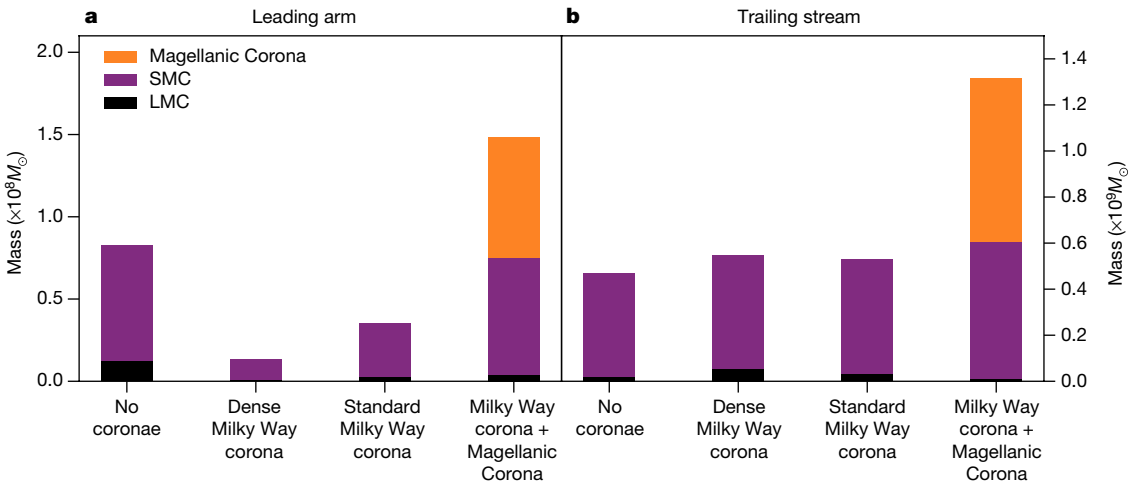


Fig. 3 | Mass budget for the Magellanic Stream. a, **b**, Origin of the mass in the leading arm (**a**) and the stream (**b**) at the present day. Each column represents a model of the formation of the stream: the fiducial dwarf–dwarf galaxy interaction model (first on the left)[10,11]; a dwarf–dwarf galaxy interaction model that includes a high-density Milky Way gas halo with total mass $5 \times 10^9 M_\odot$, which shows that the leading arm does not survive (second left column; see recent work[27]); a dwarf–dwarf galaxy interaction model that includes a lower-density Milky Way gas halo (total mass of around $2 \times 10^9 M_\odot$), still consistent with current estimates[30] (second from the right; see Extended Data Fig. 1); and the model reported here of a dwarf–dwarf galaxy interaction that includes the lower-density Milky Way gas halo and the Magellanic Corona (right-most column). The inclusion of the Magellanic Corona shows that this gas contributes greatly to the total mass of the stream, increasing it to values consistent with observations (about $1.3 \times 10^9 M_\odot$).

sensitive searches have been conducted. However, in our model, the stream is formed mostly by the warm Magellanic Corona, so its stellar counterpart is negligible. Some stars were tidally stripped from the SMC when the clouds were far from the Milky Way, but they are either phase-mixed with the Milky Way's stellar halo or extended into a thin and low-density filament of 30 mag arcsec^{-2}, which is too faint to detect with current telescopes and instrumentation.

The Magellanic Corona will be unambiguously observable via absorption in highly ionized states of carbon and silicon (C IV and Si IV) in the ultraviolet spectra of background quasars lying near the LMC on the sky. The high-ion column densities in the Magellanic Corona should decrease with increasing angular separation (impact parameter) from the LMC. In contrast to the 'down-the-barrel' studies of stars in the LMC[18,19], which pass through the interstellar medium of the LMC and may probe outflows close to the LMC disk, background-quasar sightlines offer the chance for unambiguous detections of the Magellanic Corona, because they are uncontaminated by the LMC's interstellar material.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2663-4.

1. Mathewson, D. S., Cleary, M. N. & Murray, J. D. The Magellanic stream. *Astrophys. J.* **190**, 291–296 (1974).
2. Nidever, D. L., Majewski, S. R. & Butler Burton, W. The origin of the Magellanic stream and Its leading arm. *Astrophys. J.* **679**, 432–459 (2008).
3. D'Onghia, E. & Fox, A. J. The Magellanic stream: circumnavigating the Galaxy. *Annu. Rev. Astron. Astrophys.* **54**, 363–400 (2016).
4. Brüns, C. et al. The Parkes H I survey of the Magellanic system. *Astron. Astrophys.* **432**, 45–67 (2005).
5. Kallivayalil, N., van der Marel, R. P., Besla, G., Anderson, J. & Alcock, C. Third-epoch Magellanic cloud proper motions. I. Hubble Space Telescope/WFC3 data and orbit implications. *Astrophys. J.* **764**, 161 (2013).
6. Fox, A. J. et al. The COS/UVES absorption survey of the Magellanic stream. I. One-tenth solar abundances along the body of the stream. *Astrophys. J.* **772**, 110 (2013).
7. Richter, P. et al. The COS/UVES absorption survey of the Magellanic stream. II. Evidence for a complex enrichment history of the stream from the Fairall 9 sightline. *Astrophys. J.* **772**, 111 (2013).
8. Fox, A. J. et al. The COS/UVES absorption survey of the Magellanic stream. III. Ionization, total mass, and inflow rate onto the Milky Way. *Astrophys. J.* **787**, 147 (2014).
9. Barger, K. A. et al. Revealing the ionization properties of the Magellanic stream using optical emission. *Astrophys. J.* **851**, 110 (2017).
10. Besla, G. et al. The role of dwarf galaxy interactions in shaping the Magellanic system and implications for Magellanic irregulars. *Mon. Not. R. Astron. Soc.* **421**, 2109–2138 (2012).
11. Pardy, S. A., D'Onghia, E. & Fox, A. J. Models of tidally induced gas filaments in the Magellanic stream. *Astrophys. J.* **857**, 101 (2018).
12. Hammer, F., Yang, Y. B., Flores, H., Puech, M. & Fouquet, S. The Magellanic stream system. I. Ram-pressure tails and the relics of the collision between the Magellanic clouds. *Astrophys. J.* **813**, 110 (2015).
13. Wang, J. et al. Towards a complete understanding of the Magellanic stream formation. *Mon. Not. R. Astron. Soc.* **486**, 5907–5916 (2019).
14. D'Onghia, E. & Lake, G. Small dwarf galaxies within larger dwarfs: why some are luminous while most go dark. *Astrophys. J.* **686**, L61 (2008).
15. Bechtol, K. et al. Eight new Milky Way companions discovered in first-year dark energy survey data. *Astrophys. J.* **807**, 50 (2015).
16. Nichols, M., Colless, J., Colless, M. & Bland-Hawthorn, J. Accretion of the Magellanic system onto the Galaxy. *Astrophys. J.* **742**, 110 (2011).
17. Peñarrubia, J., Gómez, F. A., Besla, G., Erkal, D. & Ma, Y.-Z. A timing constraint on the (total) mass of the Large Magellanic Cloud. *Mon. Not. R. Astron. Soc.* **456**, L54–L58 (2016).
18. Wakker, B., Howk, J. C., Chu, Y.-H., Bomans, D. & Points, S. D. Coronal C^{+3} in the Large Magellanic Cloud: evidence for a hot halo. *Astrophys. J.* **499**, L87–L91 (1998).
19. Lehner, N., Staveley-Smith, L. & Howk, J. C. Properties and origin of the high-velocity gas toward the Large Magellanic Cloud. *Astrophys. J.* **702**, 940–954 (2009).
20. Pardy, S. A. et al. Satellites of satellites: the case for Carina and Fornax. *Mon. Not. R. Astron. Soc.* **492**, 1543–1549 (2020).
21. Hafen, Z. et al. The origins of the circumgalactic medium in the FIRE simulations. *Mon. Not. R. Astron. Soc.* **488**, 1248–1272 (2019).
22. Besla, G. et al. Are the Magellanic clouds on their first passage about the Milky Way? *Astrophys. J.* **668**, 949–967 (2007).
23. Nidever, D. L., Majewski, S. R., Butler Burton, W. & Nigra, L. The 200° long Magellanic stream system. *Astrophys. J.* **723**, 1618–1631 (2010).
24. McClure-Griffiths, N. M. et al. GASS: the Parkes galactic all-sky survey. I. Survey description, goals, and initial data release. *Astrophys. J. Suppl. Ser.* **181**, 398–412 (2009).
25. Putman, M. E. et al. Tidal disruption of the Magellanic clouds by the Milky Way. *Nature* **394**, 752–754 (1998).
26. Heitsch, F. & Putman, M. E. The fate of high-velocity clouds: warm or cold cosmic rain? *Astrophys. J.* **698**, 1485–1496 (2009).
27. Tepper-García, T., Bland-Hawthorn, J., Pawlowski, M. S. & Fritz, T. K. The Magellanic system: the puzzle of the leading gas stream. *Mon. Not. R. Astron. Soc.* **488**, 918–938 (2019).
28. Nidever, D. L. et al. Spectroscopy of the young stellar association Price-Whelan 1: origin in the Magellanic leading arm and constraints on the Milky Way hot halo. *Astrophys. J.* **887**, 115 (2019).
29. Price-Whelan, A. M. et al. Discovery of a disrupting open cluster far into the Milky Way halo: a recent star formation event in the leading arm of the Magellanic stream? *Astrophys. J.* **887**, 19 (2019).
30. Bregman, J. N. et al. The extended distribution of baryons around galaxies. *Astrophys. J.* **862**, 3 (2018).

Methods

This work uses the GIZMO hydrodynamic *N*-body code[31]. GIZMO includes hydrodynamics schemes that can follow large bulk velocities and large dynamic ranges in density, making it an appropriate tool to model the hydrodynamic evolution of gas disks in isolation and when subjected to gravitational interactions. The Lagrangian meshless finite-mass method implemented in the code allows the tracking of fluid elements while capturing in detail the Kelvin–Helmholtz instabilities and shocks when the resolution is properly increased[31]. The simulations also used the adaptive gravitational softening lengths for gas particles available in GIZMO. The softening lengths are determined by the hydrodynamic smoothing lengths to ensure consistency between the gravitational and hydrodynamic calculations. These smoothing lengths are calculated using the 32 nearest neighbours for each particle. For the dark-matter component, the softening length adopted was 290 pc; for the stellar component, 100 pc was used. The simulations also implemented radiative heating and cooling[32,33], and star formation and feedback[34].

Initial set up and simulations

We created a set of *N*-body and hydrodynamic simulations of gaseous and stellar exponential disks embedded in a live NFW (Navarro–Frenk–White) dark-matter halo of Magellanic-sized galaxies[35]. The LMC progenitor galaxy has a total dark-matter-halo mass of $17.75 \times 10^{10} M_{\odot}$ ($1.8 \times 10^{5} M_{\odot}$ per particle), a stellar mass of $2.5 \times 10^{9} M_{\odot}$ ($4.2 \times 10^{3} M_{\odot}$ per particle) and a disk gas mass of $2.2 \times 10^{9} M_{\odot}$ ($4.4 \times 10^{3} M_{\odot}$ per particle). Similarly, the SMC progenitor assumes an initial total dark halo of $2.1 \times 10^{10} M_{\odot}$ ($1.9 \times 10^{5} M_{\odot}$ per particle), a stellar component of $3 \times 10^{8} M_{\odot}$ ($4.2 \times 10^{3} M_{\odot}$ per particle) and a gaseous disk of $1.6 \times 10^{9} M_{\odot}$ ($4.4 \times 10^{3} M_{\odot}$ per particle). This gives approximately 2.6×10^{6} particles in total for the Magellanic Clouds combined. For the Milky Way, a static Hernquist potential[36] was assumed, with a total mass of $10^{12} M_{\odot}$ and a scale length of 29 kpc. A live Milky Way stellar disk and bulge were also included, with masses of $4.8 \times 10^{10} M_{\odot}$ and $8 \times 10^{9} M_{\odot}$, respectively, following recent simulations[37]. The disk was included only in the full model with both coronae.

The LMC stellar disk has a scale length of 1.8 kpc, while the initial gas disk is extended with a scale length of 4.8 kpc, in agreement with isolated gaseous dwarf irregular galaxies of comparable mass[38]. Similarly, the scale length of the SMC stellar disk is initially set to 1.1 kpc and the extended gaseous disk has a scale length of 3 kpc. The outer part of the LMC disk is truncated to 25 kpc. Runs performed with the LMC outer disk truncated to various radii produce comparable results. However, for the case reported here, the filamentary structure of the trailing arm from gas tidally removed from the LMC is present, but more tenuous and less pronounced compared to previous work where the LMC disk was not truncated[11].

The Magellanic Corona is set up as a halo of warm gas surrounding the LMC, with a mass of around $3 \times 10^{9} M_{\odot}$ (about 1.5% of the LMC's total mass), extending throughout the virial radius of the LMC (roughly 100 kpc). Even though the LMC is a satellite galaxy, it is massive enough (total mass greater than $10^{11} M_{\odot}$[39,40]) to carry a group of dwarfs that includes the SMC, Carina and Fornax[20], and several additional ultrafaint dwarfs[41]. Hence, its hot corona should be at least $10^{9} M_{\odot}$ in mass[21,42]. A less massive LMC (around $5 \times 10^{10} M_{\odot}$, as inferred from the rotation curve within 8 kpc from the centre[5]) would not harbour a warm corona and would not be massive enough to carry the bright dwarfs seen in observations. Cosmological simulations confirm these estimates[43–45], and dwarf galaxies in the field have been shown to have circumgalactic gas extending out to a substantial fraction of their virial radii[46–48]. Furthermore, a conservative observational estimate of the Milky Way suggests that the circumgalactic gas is at least about 1% of the total Galactic mass. The observed mass in baryons (stars and the interstellar medium) constitutes roughly 10% of the total mass, and it is proposed that the other half of the baryons be found in the hot corona[49]. In addition, absorption-line studies show that the mass of the circumgalactic gas inside the virial radius is similar to the stellar mass[30,50]. Therefore, the total mass of the Magellanic Corona adopted here (1.5% of the LMC mass) should be considered a lower limit.

In this model, the gas properties of the Magellanic Corona surrounding the LMC are extracted from the Auriga simulations[51], a set of cosmological simulations of Milky-Way-type galaxies that contain LMC-sized satellites. The LMC analogues identified in Auriga have proper motions similar to the Hubble Space Telescope data reported for the clouds and do have an associated warm gas corona[20], the properties of which (temperature of around 5×10^{5} K, density and radial profile) are used as initial conditions for our numerical experiment. The density profile (red dashed line in Extended Data Fig. 1) decreases at larger radii, with a radial profile similar to recent results[30,52,53] for the Milky Way. The gas corona of the LMC is made up of particles with masses of $4.4 \times 10^{3} M_{\odot}$. Velocities v are assigned to gas particles according to a Maxwell–Boltzmann distribution (as in the isothermal sphere), with $f(v) \propto \exp[-mv^2/(kT)]$, where m is the mean mass per particle, k is Boltzmann's constant, and for T half the virial temperature was assumed.

At $T \approx 5 \times 10^{5}$ K, the Magellanic Corona is above the peak range of the cooling curve. Although the gaseous coronae in our models are relatively stable, owing to the inclusion of radiative heating and cooling, star formation and feedback, there may be additional physical processes included in cosmological simulations[21,51], such as feedback from active galactic nuclei, photo-ionization heating and cosmic-ray heating, that affect the stability and temperature of the circumgalactic gas[54–56].

In addition, a gas corona was set up around the Milky Way assuming an isothermal sphere of gas at $T = 1.6 \times 10^{6}$ K (the Galactic virial temperature) using the DICE code[57]. The Milky Way's gas corona does not rotate in our model, and we find that the infall of the Magellanic system does not affect the large-scale rotation of the coronal gas. As shown in previous work[27,58,59], the rotation of the Milky Way's hot corona can affect the morphology and structure of the stream; however, we are investigating the macroscale properties of the stream, which should not be affected by rotation of the Milky Way's corona. The hot corona has a total mass of around $2 \times 10^{9} M_{\odot}$, made up of particles with masses of $4.5 \times 10^{3} M_{\odot}$. It was allowed to equilibrate in isolation (with the static Milky Way dark-matter potential) for about 1 Gyr before the Magellanic Clouds fell in. The gas density profile assumed for the final run follows the distribution reported previously[30,53,60], and is displayed in Extended Data Fig. 1 (solid red line). The Magellanic Corona and the Milky Way's hot gas corona constitute an additional 2×10^{6} particles in the simulation.

Orbital parameters of the Magellanic Clouds

We carried out a parameter study of the orbital configurations of the Magellanic Clouds. Consistent with previous findings[10,11,61–64], the orbits for the LMC and SMC were set such that the clouds experience three mutual gravitational encounters before falling into the Milky Way potential. Note that the orbital configuration parameters were set to reproduce the bifurcation of the stream and the H I component, which is only 10%–20% of its total mass. In our model, the Magellanic Corona is the dominant source of the total stream mass. This result is independent of the number of encounters between the clouds and their structural parameters. The LMC orbit is obtained first by solving the differential equation of motion assuming a mass of $2 \times 10^{11} M_{\odot}$ for the LMC before the infall and a Milky Way mass of around $10^{12} M_{\odot}$. By imposing the current observed velocities and positions for the LMC as inferred from Hubble Space Telescope data, differential equations of motion are used to determine the position and velocities of the LMC at earlier times. Following previous studies[11], the SMC is initially placed 65 kpc away from the LMC on a Keplerian orbit with eccentricity $e = 0.65$ and minimum separation of 25 kpc from the LMC. The orbital history of the clouds and their mutual interactions away from the Milky Way are illustrated in Extended Data Fig. 2 and Supplementary Video 3.

After the clouds have had three close encounters, over a time period of 5.7 Gyr, the LMC and SMC are placed 220 kpc away from the centre of the Milky Way on a first pericentric passage around the Galaxy. The LMC–SMC system is rotated by 180° around the z axis, then 100° around the y axis, then −50° around the x axis. Then the LMC's centre of mass is placed at $(x, y, z) = (-22, 217, 32)$ kpc (where the Milky Way's hot corona and dark-matter potential are centred at the origin) with a velocity of $(v_x, v_y, v_z) = (18.6, -88.6, -109)$ km s^{-1}. The SMC's position and velocity were unchanged relative to the LMC for the first 5.7 Gyr in isolation. Once the clouds fall into the Milky Way, they reach their present-day positions after 1.3 Gyr, with velocities consistent with current observations[5,65]. The stream at the present day is displayed in zenithal equal-area coordinates in Extended Data Fig. 4. A fiducial model, where the stream is formed by the mutual interaction between the clouds without the inclusion of the warm and hot corona, was run first (Extended Data Fig. 4a)[11]. Subsequently, the same model assumed for the clouds was run with the inclusion of a high-density (Extended Data Fig. 4b) or low-density Milky Way hot gas corona (Extended Data Fig. 4c, d). This experiment allowed us to determine that the leading arm survives in this model if the Milky Way's gas corona has a density of $n \approx 1.7 \times 10^{-5}$ cm^{-3} at a distance of 50 kpc, in agreement with observational estimates[53] and previous studies[27]. The final run included the model of the clouds with the inclusion of both the Magellanic Corona and the Milky Way's hot halo (Extended Data Fig. 4d).

Analysis

We used a particle tracer that allows us to follow each gas particle with its temperature and density to compute the mass of the Magellanic Stream. In these numerical experiments, the stream consists of gas particles that have been stripped from the clouds and are no longer bound to the main body of their host galaxy. The gravitational potential and its kinetic energy were calculated for each gas particle. Any particle that has a larger kinetic than potential energy was considered unbound. We then projected the locations of gas particles stripped from the clouds into Magellanic coordinates and summed the masses. The gas particles were included in either the leading or trailing stream, depending on their location. We used the pygad[66] library to perform density and temperature calculations and to deposit the particles onto a mesh for visualization. The model does not include the ionization corrections to convert the hydrogen gas into the ionized fraction. The cold gas stripped from the clouds is assumed to trace the H I component, whereas the warm coronal gas is assumed to trace the ionized mass.

Data availability

The simulation data that support our findings are available at https://github.com/DOnghiaGroup/lucchini-2020-sim/. Source data are provided with this paper.

Code availability

The GIZMO code used in this work is publicly available from https://bitbucket.org/phopkins/gizmo-public/. The PyGad code used in this work is publicly available from https://bitbucket.org/broett/pygad/.

31. Hopkins, P. F. A new class of accurate, mesh-free hydrodynamic simulation methods. *Mon. Not. R. Astron. Soc.* **450**, 53–110 (2015).
32. Katz, N., Weinberg, D. H. & Hernquist, L. Cosmological simulations with TreeSPH. *Astrophys. J. Suppl. Ser.* **105**, 19–35 (1996).
33. Hopkins, P. F. et al. FIRE-2 simulations: physics versus numerics in galaxy formation. *Mon. Not. R. Astron. Soc.* **480**, 800–863 (2018).
34. Springel, V. & Hernquist, L. Cosmological smoothed particle hydrodynamics simulations: a hybrid multiphase model for star formation. *Mon. Not. R. Astron. Soc.* **339**, 289–311 (2003).
35. Springel, V. The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* **364**, 1105–1134 (2005).
36. Hernquist, L. An analytical model for spherical galaxies and bulges. *Astrophys. J.* **356**, 359–364 (1990).
37. D'Onghia, E. & Aguerri, J. A. L. Trojans in the solar neighborhood. *Astrophys. J.* **890**, 117 (2020).
38. de Blok, W. J. G. & McGaugh, S. S. The dark and visible matter content of low surface brightness disc galaxies. *Mon. Not. R. Astron. Soc.* **290**, 533–552 (1997).
39. Erkal, D. et al. Modelling the Tucana III stream - a close passage with the LMC. *Mon. Not. R. Astron. Soc.* **481**, 3148–3159 (2018).
40. Erkal, D. et al. The total mass of the Large Magellanic Cloud from its perturbation on the Orphan stream. *Mon. Not. R. Astron. Soc.* **487**, 2685–2700 (2019).
41. Kallivayalil, N. et al. The missing satellites of the Magellanic clouds? Gaia proper motions of the recently discovered ultra-faint galaxies. *Astrophys. J.* **867**, 19 (2018).
42. Jethwa, P., Erkal, D. & Belokurov, V. A Magellanic origin of the DES dwarfs. *Mon. Not. R. Astron. Soc.* **461**, 2212–2233 (2016).
43. Shen, S., Madau, P., Conroy, C., Governato, F. & Mayer, L. The baryon cycle of dwarf galaxies: dark, bursty, gas-rich polluters. *Astrophys. J.* **792**, 99 (2014).
44. Anglés-Alcázar, D. et al. The cosmic baryon cycle and galaxy mass assembly in the FIRE simulations. *Mon. Not. R. Astron. Soc.* **470**, 4698–4719 (2017).
45. Jahn, E. D. et al. Dark and luminous satellites of LMC-mass galaxies in the FIRE simulations. *Mon. Not. R. Astron. Soc.* **489**, 5348–5364 (2019).
46. Bordoloi, R. et al. The COS-dwarfs survey: the carbon reservoir around sub-L* galaxies. *Astrophys. J.* **796**, 136 (2014).
47. Johnson, S. D., Chen, H.-W., Mulchaey, J. S., Schaye, J. & Straka, L. A. The extent of chemically enriched gas around star-forming dwarf galaxies. *Astrophys. J.* **850**, L10 (2017).
48. Sokołowska, A., Mayer, L., Babul, A., Madau, P. & Shen, S. Diffuse coronae in cosmological simulations of Milky Way-sized galaxies. *Astrophys. J.* **819**, 21 (2016).
49. Fukugita, M. & Peebles, P. J. E. Massive coronae of Galaxies. *Astrophys. J.* **639**, 590–599 (2006).
50. Lehner, N. & Howk, J. C. Highly ionized plasma in the Large Magellanic Cloud: evidence for outflows and a possible galactic wind. *Mon. Not. R. Astron. Soc.* **377**, 687–704 (2007).
51. Grand, R. J. J. et al. The Auriga project: the properties and formation mechanisms of disc galaxies across cosmic time. *Mon. Not. R. Astron. Soc.* **467**, 179–207 (2017).
52. Miller, M. J. & Bregman, J. N. The structure of the Milky Way's hot gas halo. *Astrophys. J.* **770**, 118 (2013).
53. Salem, M. et al. Ram pressure stripping of the Large Magellanic Cloud's disk as a probe of the Milky Way's circumgalactic medium. *Astrophys. J.* **815**, 77 (2015).
54. Bustard, C., Pardy, S. A., D'Onghia, E., Zweibel, E. G. & Gallagher, J. S. The fate of supernova-heated gas in star-forming regions of the LMC: lessons for galaxy formation? *Astrophys. J.* **863**, 49 (2018).
55. Bustard, C., Zweibel, E. G., D'Onghia, E., Gallagher, J. S. & Farber, R. Cosmic-ray-driven outflows from the Large Magellanic Cloud: contributions to the LMC filament. *Astrophys. J.* **893**, 29 (2020).
56. Gronke, M. & Oh, S. P. How cold gas continuously entrains mass and momentum from a hot wind. *Mon. Not. R. Astron. Soc.* **492**, 1970–1990 (2020).
57. Perret, V. et al. Evolution of the mass, size, and star formation rate in high redshift merging galaxies. MIRAGE - a new sample of simulations with detailed stellar feedback. *Astron. Astrophys.* **562**, A1 (2014).
58. Bland-Hawthorn, J., Sutherland, R., Agertz, O. & Moore, B. The source of Ionization along the Magellanic stream. *Astrophys. J.* **670**, L109–L112 (2007).
59. Tepper-García, T., Bland-Hawthorn, J. & Sutherland, R. S. The Magellanic stream: break-up and accretion onto the hot Galactic corona. *Astrophys. J.* **813**, 94 (2015).
60. Faerman, Y., Sternberg, A. & McKee, C. F. Massive warm/hot galaxy coronae. II. Isentropic model. *Astrophys. J.* **893**, 82 (2020).
61. Besla, G. et al. Simulations of the Magellanic stream in a first infall scenario. *Astrophys. J.* **721**, L97–L101 (2010).
62. D'Onghia, E., Besla, G., Cox, T. J. & Hernquist, L. Resonant stripping as the origin of dwarf spheroidal galaxies. *Nature* **460**, 605–607 (2009).
63. D'Onghia, E., Vogelsberger, M., Faucher-Giguere, C.-A. & Hernquist, L. Quasi-resonant theory of tidal interactions. *Astrophys. J.* **725**, 353–368 (2010).
64. Diaz, J. D. & Bekki, K. The tidal origin of the Magellanic stream and the possibility of a stellar counterpart. *Astrophys. J.* **750**, 36 (2012).
65. Kallivayalil, N. et al. The proper motion of the Large Magellanic Cloud using HST. *Astrophys. J.* **638**, 772–785 (2006).
66. Röttgers, B. pygad: Analyzing Gadget Simulations with Python. *Astrophysics Source Code Library* 1811.014 (2018).
67. Blitz, L. & Robishaw, T. Gas-rich dwarf spheroidals. *Astrophys. J.* **541**, 675–687 (2000).
68. Stanimirović, S., Dickey, J. M., Krčo, M. & Brooks, A. M. The small-scale structure of the Magellanic stream. *Astrophys. J.* **576**, 773–789 (2002).
69. Bregman, J. N. & Lloyd-Davies, E. J. X-Ray absorption from the Milky Way halo and the local group. *Astrophys. J.* **669**, 990–1002 (2007).
70. Anderson, M. E. & Bregman, J. N. Do hot halos around galaxies contain the missing baryons? *Astrophys. J.* **714**, 320–331 (2010).
71. Murali, C. The Magellanic stream and the density of coronal gas in the Galactic halo. *Astrophys. J.* **529**, L81–L84 (2000).

Acknowledgements E.D. acknowledges the hospitality of the Center for Computational Astrophysics at the Flatiron Institute during the completion of this work.

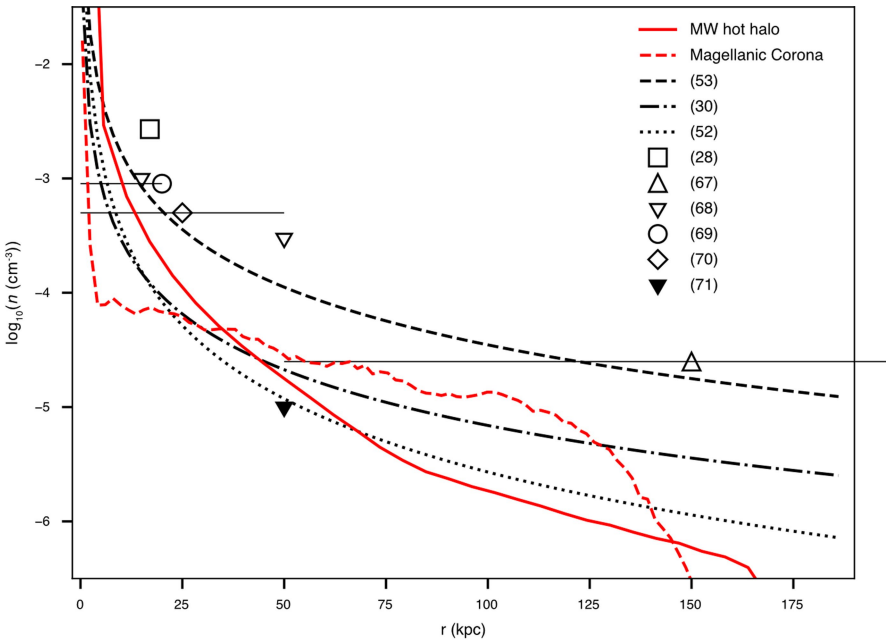
Author contributions S.L., E.D. and A.J.F. conceived and developed the numerical experiments. J.B.-H., C.B. and E.Z. contributed to discussion of the physical processes. All authors helped edit the manuscript. S.L. created the figures with input from C.B.

Competing interests The authors declare no competing interests.

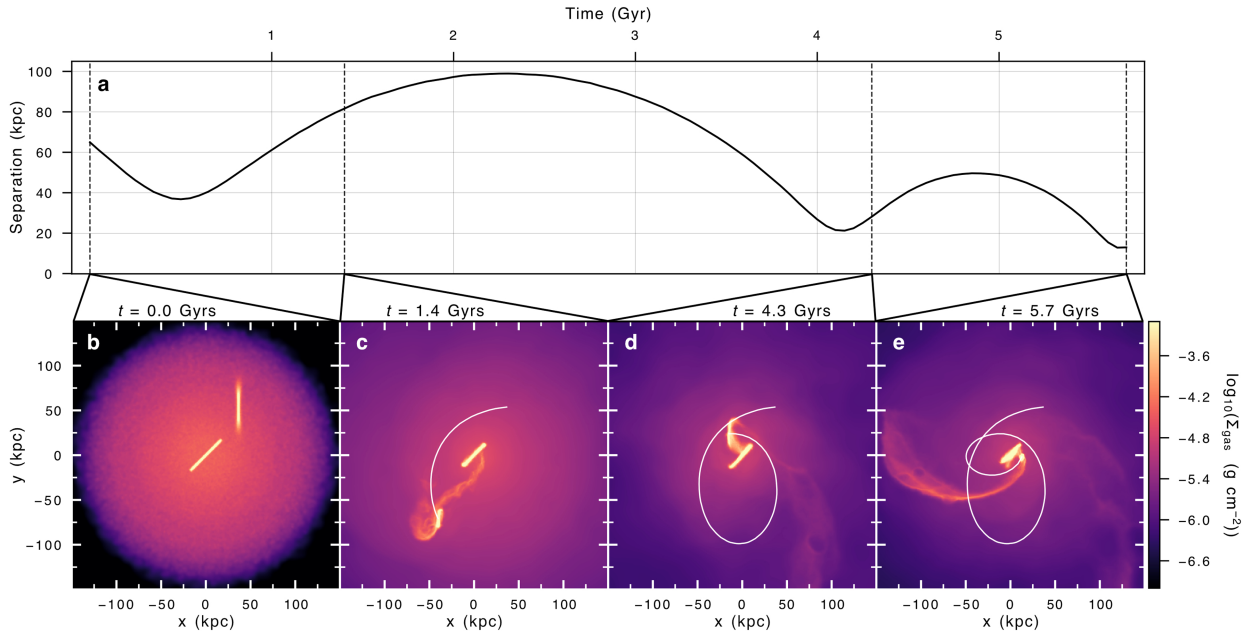
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2663-4.

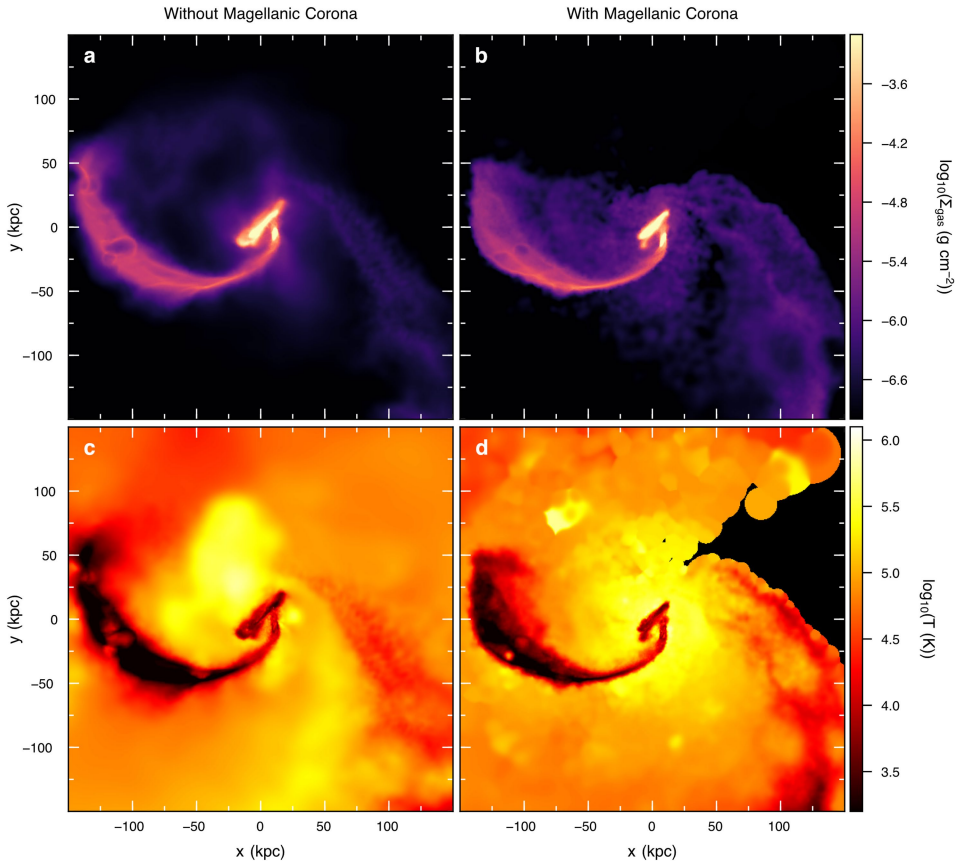
Correspondence and requests for materials should be addressed to E.D.



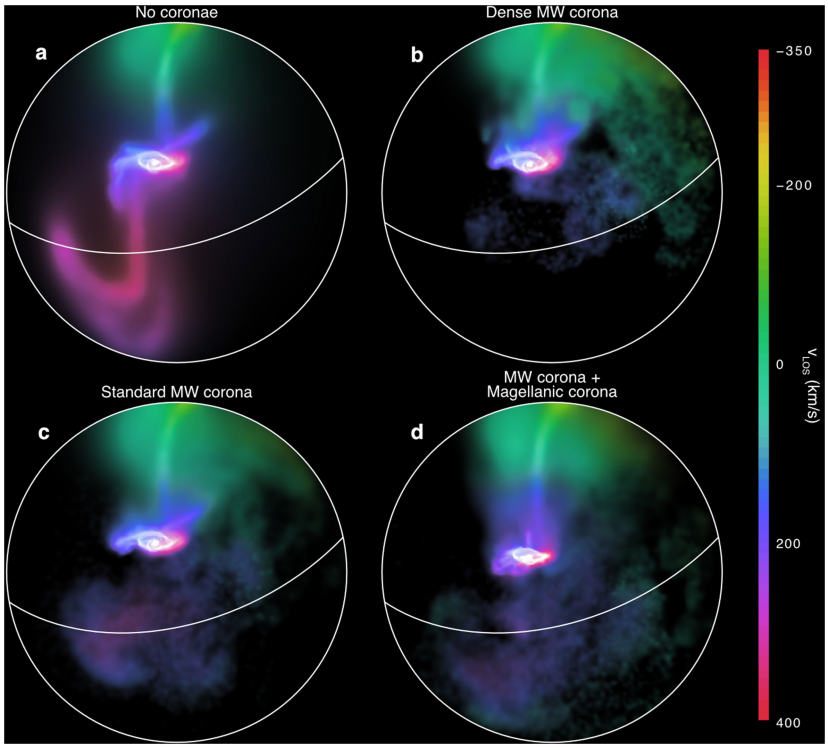
Extended Data Fig. 1 | Radial gas density profile of the Magellanic Corona and Milky Way hot corona. The number density n of gas in the models of the Magellanic Corona (dashed red line) and the Milky Way's (MW) hot corona (solid red line) is shown as a function of radius r (from the centre of the LMC and Milky Way, respectively). Estimates of the Milky Way's hot coronal density from observations are shown in black. The dotted and dot-dashed lines are fits to data[30,52,53]. The data points are labelled with the corresponding references[28,67–71], and are the same as those included in previous studies[27]. Downward (upward) pointing triangles indicate upper (lower) limits. Horizontal lines show uncertainty in radii measurements.



Extended Data Fig. 2 | Orbital histories of the LMC and SMC. a, Time evolution of the distance between the centre of mass of the LMC and the centre of mass of the SMC. The clouds interact gravitationally for a period of 5.7 Gyr (three close encounters) before falling into the Milky Way potential. **b**–**e**, Gas column density $\Sigma_{\rm gas}$ at various times during the mutual interactions between the clouds, with the orbital path of the SMC around the LMC shown as a white line (**b**, at the initial time; **c**, after 1.4 Gyr; **d**, after 4.3 Gyr; **e**, after 5.7 Gyr; marked in **a** with dotted vertical lines). The gas tidally removed from the LMC and SMC is displayed in addition to the Magellanic coronal gas.



Extended Data Fig. 3 | The effect of the Magellanic Corona on stripped gas temperature. The gas removed from the Magellanic Clouds after about 5.7 Gyr of mutual interactions (before infall into the Milky Way potential) is shown in Cartesian coordinates projected along the *z* axis onto the *x*–*y* plane. The LMC and SMC are at the centre of each panel. **a**, **b**, The gas mass surface density of the gas originating in the disks of the clouds. **c**, **d**, The gas temperature averaged along the projection axis. Results are shown for models run with (**b**, **d**) and without (**a**, **c**) the Magellanic Corona included.



Extended Data Fig. 4 | The effect of the warm and hot gas on the formation of the leading arm. The column density (brightness) and line-of-sight velocity (v_{LOS}; colour scale) for four different models for the formation of the Magellanic Stream are shown in zenithal equal-area coordinates. The white lines mark the location of the Galactic disk in the projection. These four models are the same as those shown in Fig. 3. In all four panels, only the gas originating in the gaseous disks of the Magellanic Clouds is displayed. **a**, Fiducial model, without the Milky Way's corona or Magellanic Corona (tidal forces only). **b**, A Milky Way coronal mass of $5\times10^9 M_\odot$ is included, but the Magellanic Corona is not present. The leading arm does not survive, in agreement with previous studies[27]. **c**, Same as in **b**, with the total mass of the Milky Way's hot corona reduced to $2\times10^9 M_\odot$ (see Extended Data Fig. 1), allowing the leading arm to survive. **d**, Same as in **c**, but with the addition of the Magellanic Corona. This model provides the best match to observations.

Experimental deterministic correction of qubit loss

https://doi.org/10.1038/s41586-020-2667-0

Received: 6 March 2020

Accepted: 26 June 2020

Published online: 9 September 2020

Check for updates

Roman Stricker[1]✉, Davide Vodola[2,3,4], Alexander Erhard[1], Lukas Postler[1], Michael Meth[1], Martin Ringbauer[1], Philipp Schindler[1], Thomas Monz[1,5], Markus Müller[4,6,7] & Rainer Blatt[1,8]

The successful operation of quantum computers relies on protecting qubits from decoherence and noise, which—if uncorrected—will lead to erroneous results. Because these errors accumulate during an algorithm, correcting them is a key requirement for large-scale and fault-tolerant quantum information processors. Besides computational errors, which can be addressed by quantum error correction[1–9], the carrier of the information can also be completely lost or the information can leak out of the computational space[10–14]. It is expected that such loss errors will occur at rates that are comparable to those of computational errors. Here we experimentally implement a full cycle of qubit loss detection and correction on a minimal instance of a topological surface code[15,16] in a trapped-ion quantum processor. The key technique used for this correction is a quantum non-demolition measurement performed via an ancillary qubit, which acts as a minimally invasive probe that detects absent qubits while imparting the smallest quantum mechanically possible disturbance to the remaining qubits. Upon detecting qubit loss, a recovery procedure is triggered in real time that maps the logical information onto a new encoding on the remaining qubits. Although the current demonstration is performed in a trapped-ion quantum processor[17], the protocol is applicable to other quantum computing architectures and error correcting codes, including leading two- and three-dimensional topological codes. These deterministic methods provide a complete toolbox for the correction of qubit loss that, together with techniques that mitigate computational errors, constitute the building blocks of complete and scalable quantum error correction.

Qubit loss comes in a variety of physical manifestations, such as the loss of particles encoding the qubits in atomic and photonic implementations[11–14], but also as leakage out of the two-dimensional (2D) computational qubit subspace in multi-level solid-state[18] and atomic, molecular and optical systems[11]. Whereas progress has been made in characterizing and suppressing the rate of loss and leakage processes[19–23], in many platforms these processes still occur at rates of the same order of magnitude as other errors, such as amplitude damping in trapped-ion qubits encoded in metastable states of optical transitions[11]. It is known that unnoticed and uncorrected qubit loss and leakage will severely affect the performance of quantum processors[18,24]; therefore, dedicated protocols to fight this error source have been devised. These protocols include four-qubit quantum erasure codes[10], which have been implemented using photons and post-selective quantum state analysis[12,13], as well as protocols proposed to address qubit loss in the surface code[15,25,26] and 2D colour codes[27,28]. So far, an experimental implementation of deterministic detection and correction of qubit loss and leakage, both of which will be referred to as 'loss' in the following, remains an outstanding challenge.

A general, architecture-independent protocol to protect quantum information against loss errors consists in (i) the initial encoding of logical states into a multi-qubit register, (ii) a quantum non-demolition (QND) measurement scheme that determines the position of potentially lost qubits, (iii) a reconstruction algorithm that, if not too many loss events have occurred, reconstructs the damaged code, and (iv) a final set of measurements that fixes the new code by initializing the new stabilizers.

Here, we encode a single logical qubit in an excerpt of the surface code[15,16], which is a topological quantum error-correcting (QEC) code in which physical qubits reside on the edges of a 2D square lattice; see Fig. 1a. The surface code is a Calderbank–Shor–Steane code[29,30], for which stabilizer operators S_V^X are associated to each vertex V (blue cross in Fig. 1a) via $S_V^X = \prod_{j\in V} X_j$ and to each plaquette P (green square in Fig. 1a) via $S_P^Z = \prod_{j\in P} Z_j$, where X_j, Y_j, Z_j are Pauli matrices acting on the physical qubit j. All stabilizers mutually commute, and their common +1 eigenspace fixes the code space that hosts the logical quantum states $|\psi_L\rangle$, that is, $S_P^Z|\psi_L\rangle = S_V^X|\psi_L\rangle = |\psi_L\rangle$ for all plaquettes and vertices. The operators that define and induce flips of the logical-basis states

[1]Institut für Experimentalphysik, Universität Innsbruck, Innsbruck, Austria. [2]Dipartimento di Fisica e Astronomia dell'Università di Bologna, Bologna, Italy. [3]INFN, Sezione di Bologna, Bologna, Italy. [4]Department of Physics, College of Science, Swansea University, Swansea, UK. [5]Alpine Quantum Technologies GmbH, Innsbruck, Austria. [6]Institute for Quantum Information, RWTH Aachen University, Aachen, Germany. [7]Theoretical Nanoelectronics, Peter Grünberg Institute, Forschungszentrum Jülich, Jülich, Germany. [8]Institut für Quantenoptik und Quanteninformation, Österreichische Akademie der Wissenschaften, Innsbruck, Austria. ✉e-mail: roman.stricker@uibk.ac.at

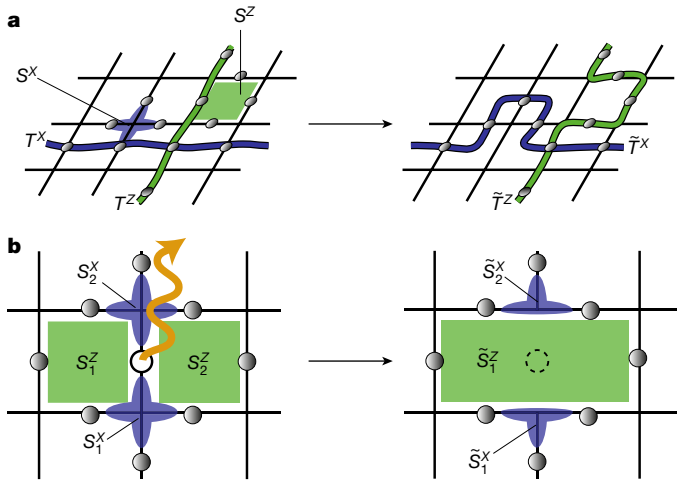


Fig. 1 | The surface code and correction of qubit loss. a, Logical qubits are encoded collectively in many physical qubits (grey circles) that are located on the edges of a 2D square lattice. The code space is defined via four-qubit S^Z and S^X stabilizers acting on groups of qubits that reside around plaquettes (green square) and vertices (blue cross) of the lattice. Logical T^Z and T^X operators are defined along strings of qubits that span the entire lattice along two non-trivial paths, as depicted by the vertical green (horizontal blue) string for T^Z (T^X). Right, logical string operators do not have unique support, but can be deformed by multiplication with stabilizers, as illustrated for T^Z (T^X), which is deformed into $\tilde{T}^Z$ ($\tilde{T}^X$) by the green plaquette (blue vertex) stabilizer. **b**, Left, excerpt of a qubit lattice suffering the loss (orange arrow) of a physical qubit (white circle). The loss affects two plaquette operators, S_1^Z and S_2^Z, and two vertex operators, S_1^X and S_2^X. Right, the correction algorithm consists of introducing a new merged Z-stabilizer generator as $\tilde{S}_1^Z = S_1^Z S_2^Z$, which does not involve the lost qubit, and two new X stabilizers, $\tilde{S}_1^X$ and $\tilde{S}_2^X$, which have reduced support on three qubits that are unaffected by the loss.

$|0_L\rangle$ and $|1_L\rangle$ are the logical generators T^Z and T^X, respectively. They commute with all stabilizers, and can be chosen as products of X and Z operators along strings that span the entire lattice; see Fig. 1a.

To recover a logical qubit affected by qubit loss, one needs to switch to an equivalent set of stabilizers $\{\tilde{S}_V^X, \tilde{S}_P^Z\}$ and logical operators $\{\tilde{T}^X, \tilde{T}^Z\}$ defined only on qubits that are not affected by losses. For this redefinition we follow the scheme introduced in ref. [25] and shown in Fig. 1b. Notably, the logical operators do not have unique support because equivalent operators $\tilde{T}^X$ and $\tilde{T}^Z$ can be obtained by multiplying T^X and T^Z by any subset of stabilizers. For the surface code, this results in the deformation of the string of physical qubits that supports the logical operator; see Fig. 1a. For too many losses, however, finding such an equivalent logical operator might not be possible. Because each loss event results in the deletion of one edge (bond) of the 2D square lattice, the question of whether such a path supporting a logical operator exists corresponds to the classical problem of bond percolation, which for the surface code results in a threshold of tolerable qubit loss rate as high as 50% in the absence of other errors[25].

Inspired by the surface code stabilizer structure, we implement a subspace defined by three stabilizers on four qubits that allows us to experimentally explore the reconstruction protocol as described in Fig. 2a. We note that this subspace is neither an error detection nor a correction code for Pauli errors, but the logical information can be made robust to the loss of qubit 1. For the physical realization of this code, we consider a string of $^{40}Ca^+$ ions confined in a linear Paul trap[17]. Each ion represents a physical qubit encoded in the electronic levels $S_{1/2}(m=-1/2) = |0\rangle$ and $D_{5/2}(m=-1/2) = |1\rangle$. Our setup is capable of realizing a universal set of quantum gate operations consisting in (a) single-qubit rotations by an angle θ around the z axis of the form $R_j^Z(\theta) = \exp(-(i\theta Z_j)/2)$ on the jth ion, (b) collective qubit rotations around the x and y axes of the form $R^\sigma(\theta) = \exp(-i\theta \sum_j (\sigma_j/2))$, with $\sigma = X$ or Y, via a laser beam addressing the entire register, and (c) multi-qubit Mølmer–Sørensen entangling gate operations[31] $\mathrm{MS}^X(\theta) = \exp(-i\theta \sum_{j<\ell} (X_j X_\ell/2))$. This gate set is complemented by single-qubit hiding and unhiding operations in order to apply collective multi-qubit operations to only a subset of qubits[17]. Similarly, this technique is used to read out individual qubits within the register without influencing the other qubits; see Supplementary Information for details.

To benchmark the performance of the protocol we introduce qubit loss in a controlled way as leakage to another electronic level outside the computational subspace; see Fig. 3a. Leakage is the dominant form of qubit loss in ion-trap architectures, whereas our protocol is also applicable to other forms of loss and architectures. The qubit that potentially suffers a loss is partially pumped out of its computational subspace $\{S_{1/2}(m=-1/2) = |0\rangle, D_{5/2}(m=-1/2) = |1\rangle\}$ by coherently driving the carrier transition $S_{1/2}(m=-1/2) = |0\rangle \leftrightarrow D_{5/2}(m=-5/2) = |2\rangle$. In the following, this is referred to as the loss operation $R_{\mathrm{loss}}(\phi)$, where the probability of loss from state $|0\rangle$ is given by $\sin^2(\phi/2)$. The loss rate on the logical qubit is the product of the loss probability with the population in $|0\rangle$.

To detect a loss event we implement a QND measurement as shown in Fig. 3a, which signals the loss of a code qubit by a bit-flip on an ancillary qubit prepared in state $|0\rangle$, followed by an addressed readout of the ancillary qubit. The key ingredient of this QND measurement is a two-qubit entangling gate operation $\mathrm{MS}^X(\pi)$ that performs a collective bit-flip operation on the code and ancilla qubits if the code qubit is present. If the code qubit has been lost, on the other hand, regardless of whether loss occurs from the $|0\rangle$ or $|1\rangle$ state, this operation acts only on the ancilla, on which it performs an identity operation; see Supplementary Information for details. A subsequent collective bit-flip $R^X(\pi) = X$ will flip the ancilla qubit to $|1\rangle$ before its addressed readout. If no loss occurred, the collective bit-flip induced by $\mathrm{MS}^X(\pi)$ will be undone by the $R^X(\pi) = X$ operation, and the ancilla qubit will end in state $|0\rangle$ (ref. [17]). The code qubit, on the other hand, will in this case undergo a non-unitary evolution given by (up to normalization) $\rho \to E\rho E^\dagger$ with $E = |1\rangle\langle 1| + \cos(\phi/2)|0\rangle\langle 0|$, which for small loss rates ($\phi \approx 0$) converges to the identity operation. This is a consequence of the information gain that no loss has occurred in this instance, provided by the ancilla measurement; see Supplementary Information.

We test the loss-detection sub-circuit on the full five-qubit register by driving the loss transition $R_{\mathrm{loss}}(\phi)$ on qubit 1 and measuring the population in the $D_{5/2}$ state on both the code and ancilla qubits. This measurement does not distinguish between the different Zeeman sublevels of the $D_{5/2}$-state manifold. Figure 3b shows that loss detected by the ancilla qubit matches the loss induced on qubit 1 within statistical uncertainty, indicating that a loss event is reliably detected. The quantified detection efficiency is 96.5(4)%, with a false positive rate of 3(1)% and a false negative rate of 1(1)%.

We note that for very low loss rates, the fidelity of the final state after correcting qubit loss will be limited by imperfections in the QND loss-detection unit; see Supplementary Information for details. To quantify the performance of the QND detection scheme in the absence of loss, we reconstruct the Choi matrix[32] of the corresponding non-unitary map using generalized quantum process tomography. The reconstructed Choi matrix shown in Fig. 3c confirms this dynamical behaviour expected in the no-loss case with a process fidelity of 90(2)% with ~20% ($\phi = 0.3\pi$) loss from $|0\rangle$. This demonstrates that information about loss on the code qubit can be reliably mapped onto the ancilla qubit. For general loss-detection purposes, one could use the detection unit to probe all code qubits within the register sequentially.

To investigate the robustness of our minimal-instance logical qubit against loss, we combine the loss-detection unit and the conditional-correction step in a 1+4-qubit algorithm, sketched in Fig. 2a. The experimental sequence for encoding an arbitrary input state of the form $|\psi_L\rangle = \cos(\alpha/2)|0_L\rangle + i\sin(\alpha/2)|1_L\rangle$ in our ion-trap quantum computer is given in Supplementary Information. The logical basis states $|0_L\rangle$ and $|1_L\rangle$ encoded by the initial stabilizers read

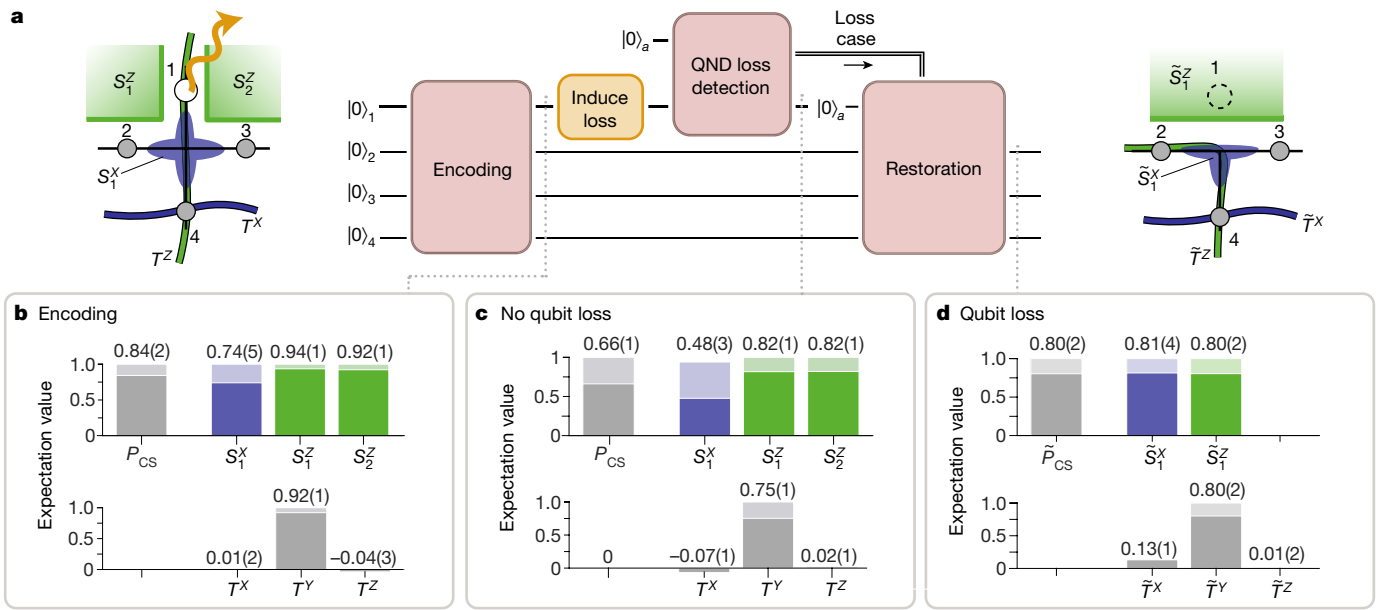


Fig. 2 | Experimental realization of the 1+4-qubit algorithm aiming at loss detection and correction. a, Minimal four-qubit system for the experimental realization of the full loss-correction protocol. The code is defined by three stabilizers, $S_1^Z = Z_1Z_2$, $S_2^Z = Z_1Z_3$ (green squares) and $S_1^X = X_1X_2X_3X_4$ (blue cross) and stores a single logical qubit with logical operators $T^Z = Z_1Z_4$, $T^X = X_4$ and $T^Y = iT^XT^Z$. In the event of the loss (orange arrow) of qubit 1 (white circle), the merged Z stabilizer $\tilde{S}_1^Z = S_1^ZS_2^Z = Z_2Z_3$ and a new X stabilizer $\tilde{S}_1^X = X_2X_3X_4$ with reduced support on the remaining three qubits are introduced for the new encoding. The logical operators equivalent to the initial ones are $\tilde{T}^Z = S_1^ZT^Z = Z_2Z_4$, $\tilde{T}^X = X_4$ and $\tilde{T}^Y = i\tilde{T}^X\tilde{T}^Z$. **b**, Expectation values for logical operators (T), stabilizers (S) and code space populations (P_{CS}), defined in Supplementary Information, for the logical superposition state $|+i_L\rangle = (|0_L\rangle + i|1_L\rangle)/\sqrt{2}$. A loss rate of 25% was induced on qubit 1. All values are estimated from four-qubit quantum state tomography, with ideal values shaded in the background. Errors correspond to one standard deviation of statistical uncertainty due to quantum projection noise. **c**, In the absence of loss, the logical encoding remains largely intact. **d**, In the case of loss, we reconstruct the code on the three remaining qubits after measuring the shrunk stabilizer of the new encoding, $\tilde{S}_1^X = X_2X_3X_4$, and selecting the appropriate Pauli basis, that is, performing a Pauli frame update in the case of a −1 outcome in the $\tilde{S}_1^X$ measurement.

$|0_L\rangle = (|0000\rangle + |1111\rangle)/\sqrt{2}$ and $|1_L\rangle = (|0001\rangle + |1110\rangle)/\sqrt{2}$. These entangled states are produced with a single fully entangling MS gate, $MS^X(\pi/2)$, acting on all four code qubits, supported by additional local operations. Loss is observed using the QND detection unit, with an ancilla qubit for loss readout. In this smallest excerpt of the surface code, we consider potential qubit loss to happen on qubit 1 only; hence, we probe only qubit 1 using the QND-detection unit as indicated in Fig. 2a. Conditional on the detection of a loss event, our control scheme triggers a real-time deterministic code restoration via feed-forward. If no loss is detected, the logical states can be verified by measuring the generators of the stabilizer group $\{S_1^Z = Z_1Z_2, S_2^Z = Z_1Z_3, S_1^X = X_1X_2X_3X_4\}$ and the logical operators $\{T^Z = Z_1Z_4, T^X = X_4, T^Y = iT^XT^Z\}$ of the original encoding. If loss occurs, the encoded logical information can be restored by switching to an encoding defined on a smaller subset of three qubits. This is realized by a projective measurement of the shrunk stabilizer of the new encoding $\tilde{S}_1^X = X_2X_3X_4$, which after the loss is in an undetermined state. This

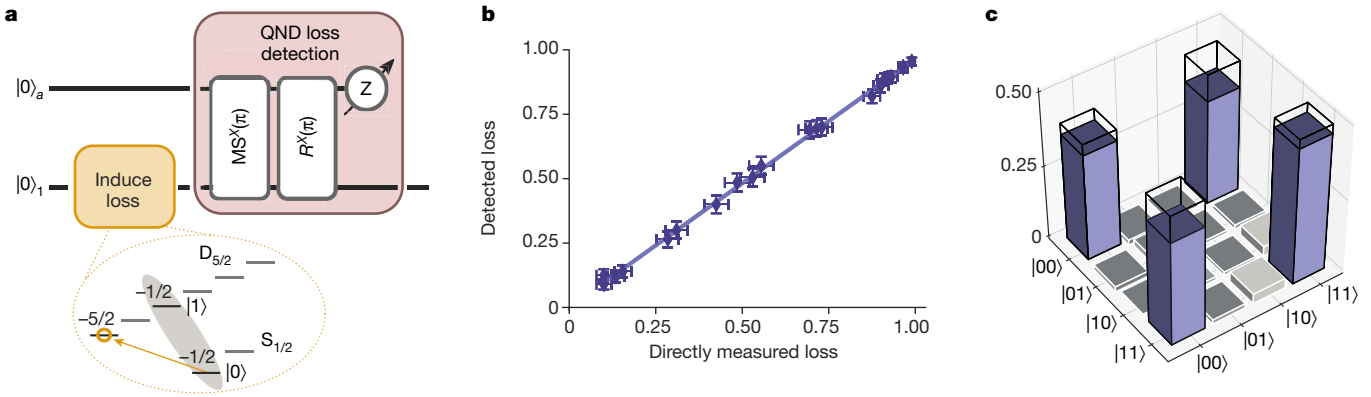


Fig. 3 | Investigating the performance of the QND loss-detection unit. a, Circuit representation of the detection unit, which maps potential loss from qubit 1 onto the ancilla qubit. The experimental results in **b** and **c** were extracted from experiments performed on the full five-qubit register, according to Fig. 2. **b**, Population in the $D_{5/2}$ state of qubit 1 (directly measured loss) and ancilla qubit (detected loss) measured after loss detection. Controlled loss of up to 100% from state $|0\rangle$ was introduced. The estimated detection efficiency is 96.5(4)%, which demonstrates that the occurrence of a loss event can be reliably mapped onto the ancilla qubit and read out in a QND fashion. Errors correspond to one standard deviation of statistical uncertainty due to quantum projection noise. **c**, Reconstructed Choi matrix for a loss of ~20% ($\phi = 0.3\pi$) from the $|0\rangle$ state with a process fidelity of 90(2)%, compared to the ideal values denoted by black frames. We find that, as expected, the detection unit performs a non-unitary evolution that deviates from the identity operator owing to measurement back-action; see Supplementary Information.

initializes the three-qubit stabilizer in a +1 (or −1) eigenstate, where the −1 case requires a redefinition of the Pauli basis (Pauli frame update)[33,34]; see Supplementary Information for details. For this stabilizer readout, a freshly initialized ancilla qubit is needed. In our implementation we recycle the ancilla qubit, previously used for the QND loss detection, because it remains unaffected by the measurement in the loss case. Following this procedure, the initial logical encoding is reconstructed in the smaller subset of three qubits; see Fig. 2a.

We now present the results obtained from the full implementation of the 1+4-qubit algorithm, as shown in Fig. 2. Data were taken for three different input states, namely, the logical basis states $|0_L\rangle$ and $|1_L\rangle$, presented in Supplementary Information, as well as their superposition $|+i_L\rangle = (|0_L\rangle + i|1_L\rangle)/\sqrt{2}$ presented here. To verify the initialization of $|+i_L\rangle$, we reconstruct the experimental density matrix via four-qubit quantum state tomography on the code qubits, yielding a fidelity of 84(1)% with the ideal state. From the reconstructed density matrix we further extract the components of the 'logical' Bloch vector, represented by expectation values of the associated logical operators, the code space population P_{CS} (explained in the Supplementary Information) and the expectation values of the stabilizer generators summarized in Fig. 2b.

After the encoding, partial loss on qubit 1 is induced by coherently exciting the loss transition $R_{loss}(\phi)$ for different values of ϕ. Here, we present the case of a loss rate of 25%, that is, $\phi = 0.5\pi$, and other values are found in Supplementary Information. Loss is detected by a QND measurement mapping the information of loss onto the ancilla qubit, followed by a projective measurement of the ancilla qubit. The measurement result triggers a real-time deterministic code restoration via feed-forward. If no loss is detected, quantum state tomography on all four code qubits is performed to verify that the initial encoding $|+i_L\rangle$ is still intact, with a fidelity of 66(1)% with respect to the expected state; see Fig. 2c. If loss is detected, the code is switched to the remaining three qubits by a projective measurement of the shrunk stabilizer $\tilde{S}_1^X$, as illustrated in Fig. 2a, and a Pauli frame update in case of a −1 outcome. Quantum state tomography yields a fidelity of the resulting three-qubit logical state $|+i_L\rangle$ of 78(1)%; see Fig. 2d.

The observed decrease in fidelity after loss detection is mainly due to cross-talk between neighbouring ions resulting in unitary errors on the final state, and dephasing due to laser-frequency and magnetic-field fluctuations. Additionally, in the no-loss case the ancilla qubit has scattered photons during the in-sequence loss detection. This heats up the ion string, decreasing the quality of the subsequent tomography operations.

Our work demonstrates the first deterministic detection and correction of qubit loss. Our building blocks are readily applicable to leading QEC codes, such as the surface and colour codes, and fully compatible with the framework of topological QEC. Although this demonstration is performed on an ion quantum processor, essentially all experimental quantum computing platforms are affected by qubit loss or leakage, and could thus benefit from our methods. Fault-tolerant versions of the presented routines in combination with correction of computational errors constitute required extensions towards the realization of large-scale quantum computers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2667-0.

1. Gottesman, D. Theory of fault-tolerant quantum computation. *Phys. Rev. A* **57**, 127–137 (1998).
2. Chiaverini, J. et al. Realization of quantum error correction. *Nature* **432**, 602–605 (2004).
3. Schindler, P. et al. Experimental repetitive quantum error correction. *Science* **332**, 1059–1061 (2011).
4. Nigg, D. et al. Quantum computations on a topologically encoded qubit. *Science* **345**, 302–305 (2014).
5. Takita, M., Cross, A. W., Córcoles, A. D., Chow, J. M. & Gambetta, J. M. Experimental demonstration of fault-tolerant state preparation with superconducting qubits. *Phys. Rev. Lett.* **119**, 180501 (2017).
6. Linke, N. M. et al. Fault-tolerant quantum error detection. *Sci. Adv.* **3**, e1701074 (2017).
7. Córcoles, A. D. et al. Demonstration of a quantum error detection code using a square lattice of four superconducting qubits. *Nat. Commun.* **6**, 6979 (2015).
8. Knill, E., Laflamme, R., Martinez, R. & Negrevergne, C. Benchmarking quantum computers: the five-qubit error correcting code. *Phys. Rev. Lett.* **86**, 5811 (2001).
9. Yao, X.-C. et al. Experimental demonstration of topological error correction. *Nature* **482**, 489 (2012).
10. Grassl, M., Beth, T. & Pellizzari, T. Codes for the quantum erasure channel. *Phys. Rev. A* **56**, 33–38 (1997).
11. Brown, N. C. & Brown, K. R. Comparing Zeeman qubits to hyperfine qubits in the context of the surface code: $^{174}Yb^+$ and $^{171}Yb^+$. *Phys. Rev. A* **97**, 052301 (2018).
12. Lu, C.-Y. et al. Experimental quantum coding against qubit loss error. *Proc. Natl Acad. Sci. USA* **105**, 11050–11054 (2008).
13. Bell, B. A. et al. Experimental demonstration of a graph state quantum error-correction code. *Nat. Commun.* **5**, 3658 (2014).
14. Morley-Short, S. et al. Physical-depth architectural requirements for generating universal photonic cluster states. *Quant. Sci. Tech.* **3**, 015005 (2018).
15. Kitaev, A. Fault-tolerant quantum computation by anyons. *Ann. Phys.* **303**, 2–30 (2003).
16. Dennis, E., Kitaev, A., Landahl, A. & Preskill J. Topological quantum memory. *J. Math. Phys.* **43**, 4452 (2002).
17. Schindler, P. et al. A quantum information processor with trapped ions. *New J. Phys.* **15**, 123012 (2013).
18. Fowler, A. G. Coping with qubit leakage in topological codes. *Phys. Rev. A* **88**, 042308 (2013).
19. Epstein, J. M., Cross, A. W., Magesan, E. & Gambetta, J. M. Investigating the limits of randomized benchmarking protocols. *Phys. Rev. A* **89**, 062321 (2014).
20. Xia, T. et al. Randomized benchmarking of single-qubit gates in a 2D array of neutral-atom qubits. *Phys. Rev. Lett.* **114**, 100503 (2015).
21. Kwon, M., Ebert, M. F., Walker, T. G. & Saffman, M. Parallel low-loss measurement of multiple atomic qubits. *Phys. Rev. Lett.* **119**, 180504 (2017).
22. Brown, N. C. & Brown, K. R. Leakage mitigation for quantum error correction using a mixed qubit scheme. *Phys. Rev. A* **100**, 032325 (2019).
23. Hayes, D. et al. Eliminating leakage errors in hyperfine qubits. *Phys. Rev. Lett.* **124**, 170501 (2020).
24. Ghosh, J., Fowler, A. G., Martinis, J. M. & Geller, M. R. Understanding the effects of leakage in superconducting quantum-error-detection circuits. *Phys. Rev. A* **88**, 062329 (2013).
25. Stace, T. M., Barrett, S. D. & Doherty, A. C. Thresholds for topological codes in the presence of loss. *Phys. Rev. Lett.* **102**, 200501 (2009).
26. Varbanov, B. M. et al. Leakage detection for a transmon-based surface code. Preprint at https://arxiv.org/abs/2002.07119 (2020).
27. Bombin, H. & Martin-Delgado, M. A. Topological quantum distillation. *Phys. Rev. Lett.* **97**, 180501 (2006).
28. Vodola, D., Amaro, D., Martin-Delgado, M. A. & Müller, M. Twins percolation for qubit losses in topological color codes. *Phys. Rev. Lett.* **121**, 060501 (2018).
29. Calderbank, A. R. & Shor, P. W. Good quantum error-correcting codes exist. *Phys. Rev. A* **54**, 1098–1105 (1996).
30. Steane, A. M. Error correcting codes in quantum theory. *Phys. Rev. Lett.* **77**, 793–797 (1996).
31. Mølmer, K. & Sørensen, A. Multiparticle entanglement of hot trapped ions. *Phys. Rev. Lett.* **82**, 1835–1838 (1999).
32. Choi, M.-D. Completely positive linear maps on complex matrices. *Linear Algebra Appl.* **10**, 285–290 (1975).
33. Knill, E. Quantum computing with realistically noisy devices. *Nature* **434**, 39–44 (2005).
34. Aliferis, P., Gottesman, D. & Preskill, J. Quantum accuracy threshold for concatenated distance-3 codes. *Quantum Inf. Comput.* **6**, 97–165 (2006).

Data availability

The data underlying the findings of this work are available at https://doi.org/10.5281/zenodo.3900057. Source data are provided with this paper.

Code availability

All codes used for data analysis are available from the corresponding author upon reasonable request.

Acknowledgements We gratefully acknowledge funding by the US Army Research Office (ARO) through grant number W911NF-14-1-0103. We also acknowledge funding by the Austrian Science Fund (FWF), through the SFB BeyondC (FWF Project number F71), by the Austrian Research Promotion Agency (FFG) contract 872766, by the EU H2020-FETFLAG-2018-03 under Grant Agreement number 820495, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via US ARO Grant number W911NF-16-1-0070. All statements of fact, opinions or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of ODNI, the IARPA, or the US Government. We acknowledge support from the Samsung Advanced Institute of Technology Global Research Outreach. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement number 801110 and the Austrian Federal Ministry of Education, Science and Research (BMBWF). The information provided in this Article reflects only the authors' views; the EU Agency is not responsible for any use that may be made of this information.

Author contributions D.V. and M. Müller derived the theory results. R.S., A.E., L.P., M. Meth, M.R., P.S. and T.M. performed the experiments. R.S. analysed the data. T.M., M. Müller and R.B. supervised the project. All authors contributed to the writing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2667-0.
Correspondence and requests for materials should be addressed to R.S.
Peer review information *Nature* thanks Tom Stace and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Co-designing electronics with microfluidics for more sustainable cooling

https://doi.org/10.1038/s41586-020-2666-1

Received: 4 March 2020

Accepted: 10 July 2020

Published online: 9 September 2020

Check for updates

Remco van Erp[1], Reza Soleimanzadeh[1], Luca Nela[1], Georgios Kampitsis[1] & Elison Matioli[1]✉

Thermal management is one of the main challenges for the future of electronics[1–5]. With the ever-increasing rate of data generation and communication, as well as the constant push to reduce the size and costs of industrial converter systems, the power density of electronics has risen[6]. Consequently, cooling, with its enormous energy and water consumption, has an increasingly large environmental impact[7,8], and new technologies are needed to extract the heat in a more sustainable way—that is, requiring less water and energy[9]. Embedding liquid cooling directly inside the chip is a promising approach for more efficient thermal management[5,10,11]. However, even in state-of-the-art approaches, the electronics and cooling are treated separately, leaving the full energy-saving potential of embedded cooling untapped. Here we show that by co-designing microfluidics and electronics within the same semiconductor substrate we can produce a monolithically integrated manifold microchannel cooling structure with efficiency beyond what is currently available. Our results show that heat fluxes exceeding 1.7 kilowatts per square centimetre can be extracted using only 0.57 watts per square centimetre of pumping power. We observed an unprecedented coefficient of performance (exceeding 10,000) for single-phase water-cooling of heat fluxes exceeding 1 kilowatt per square centimetre, corresponding to a 50-fold increase compared to straight microchannels, as well as a very high average Nusselt number of 16. The proposed cooling technology should enable further miniaturization of electronics, potentially extending Moore's law and greatly reducing the energy consumption in cooling of electronics. Furthermore, by removing the need for large external heat sinks, this approach should enable the realization of very compact power converters integrated on a single chip.

In the USA alone, data centres consume 24 TWh of electricity and 100 billion litres of water to satisfy their cooling demands[8], corresponding to the residential needs of a city of the size of Philadelphia[12–14]. The environmental impact of this information technology infrastructure is expected to increase dramatically[9], for example, accounting for up to 31% of Ireland's electricity demand by 2027 (ref. [15]), in large part due to the power consumption of cooling systems. This development is accompanied by the constant push to shrink the size of semiconductor devices, which results in higher heat fluxes that become increasingly challenging to extract and require new cooling solutions. A similar need is observed in power electronics, as the electrification of our society demands more powerful, more efficient and smaller energy conversion systems. Wide-bandgap semiconductors, such as gallium nitride (GaN), are promising candidates for this purpose[16]. These materials enable much smaller dies than those of traditional semiconductors as well as the monolithic integration of power devices, supporting the miniaturization of complete power converters into a single chip[17]. However, to unlock the full potential of GaN, strategies for sustainable cooling of high-heat-flux applications are required.

Substantial research efforts have focused on improving the thermal path between the hotspot and the coolant. However, heat extraction capability is fundamentally limited by the thermal resistance between the semiconductor die and packaging. Furthermore, relying on large heat sinks reduces the power density and hinders integration, since devices cannot be densely packed. Bringing the coolant in direct contact with the device may be a way to overcome this limiting factor, for example, by impinging coolant on a bare die[18] or by etching micrometre-sized channels directly inside the device to turn the substrate into a heat sink. The latter technique demonstrated state-of-the-art cooling performance due to the highly efficient heat transfer at the microscale[19–21]. The high pressure drop and large temperature gradients associated with these straight, parallel microchannels (SPMCs) were overcome by splitting the flow into multiple parallel sections, and distributing the coolant over these channels using manifolds[22]. Early investigations[23–26] and systematic numerical studies[27–30] of manifold microchannel (MMC) heat sinks showed a large reduction in pumping power requirements and thermal resistance compared to SPMCs. Excellent heat extraction has been demonstrated with copper MMCs[31], compact micro-fabricated multilayer silicon structures[32–35] and by using additive manufacturing[36,37]. However, in all these approaches, the heat sink and electronic structure and fabrication process are considered separately, either by integrating a simple resistive heater functioning as the heat source, or

[1]Power and Wide-band-gap Electronics Research Laboratory (POWERlab), Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
✉e-mail: elison.matioli@epfl.ch

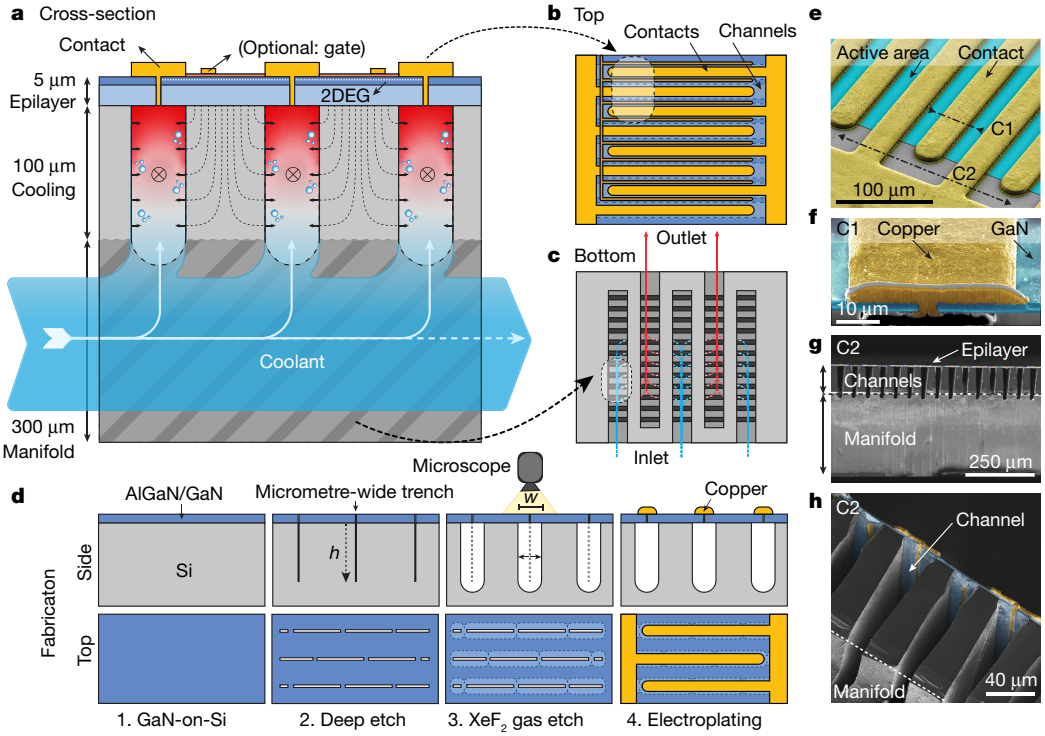


Fig. 1 | Co-designed microfluidically cooled electric device. a, Schematic of the device structure, in which the AlGaN/GaN epilayer provides the electronic functions and the silicon functions as cooling and fluid distribution manifold. Metal contacts seal the buried microchannels embedded underneath. Coolant coming from the manifolds flows in the out-of-plane orientation inside the microchannels to remove heat from the device. **b**, Top view of the co-designed device structure: each contact is aligned and seals the buried channel in a scaled-up multi-finger structure. **c**, Bottom view of the co-designed device structure: the manifold structure distributes the flow over the microchannels. **d**, Summary of the proposed cooling method: a staggered pattern of narrow high-aspect-ratio slits is first etched through the AlGaN/GaN epilayer into the silicon. Next, an isotropic gas etch widens the channels in the silicon, coalescing under the epilayer. The openings in the epilayer are then sealed using electroplating. **e**, SEM image of the AlGaN/GaN surface after sealing the microchannels. Contact pads hermetically seal the incisions in the AlGaN/GaN epilayer. **f**, Cross-sectional SEM image along C1, showing the incision in the epilayer sealed with electroplated copper. **g**, Cross-sectional SEM image along C2, showing an array of buried microchannels, as well as a sidewall of the perpendicular manifold channel. **h**, Close-up of the cross-sectional image along C2, showing the exposed microchannel below the electroplated-copper sealing layer.

by bonding the MMC structure to a commercial device[38]. This leaves the large potential of MMCs untapped. Improving the thermal coupling between the heat source and cooling has been investigated for hotspot mitigation[39–41], but has remained unexplored in a complete device structure. Furthermore, despite much MMC heat sink research, the increasing complexity and associated reliability concerns caused by the multiple bonded layers required for coolant delivery have prevented the adoption of MMCs in commercial devices.

In this work, we address these concerns by combining cooling and device design, using an approach in which a MMC heat sink is designed and fabricated in conjunction with the electronics. We present a monolithically integrated manifold microchannel (mMMC) heat sink in a single-crystalline silicon substrate with an epilayer, produced without the need for cumbersome bonding steps. Here the device design and heat-sink fabrication are combined within the same process, with buried cooling channels embedded directly below the active area of the chip. Coolant thus impinging directly on the heat sources provides local and efficient heat extraction (Fig. 1a). On the back of this same substrate, manifold channels spread the liquid over the die (Fig. 1c) to obtain high temperature uniformity and low pressure drop, leading to a very low pumping-power consumption and vastly improved cooling performance. Since the electronics and microfluidics are fully coupled and aligned (Fig. 1b), we call this approach microfluidic–electronic co-design. We demonstrated microfluidic–electronic co-design on GaN-on-Si, a low-cost platform that is promising for realizing high-power converters on a chip, comprising a GaN epilayer a few micrometres thick on a low-cost silicon substrate. The passive silicon substrate typically lacks functionality, but by turning it into an active cooling layer, it has the potential to extract extreme heat fluxes, without requiring the added cost of high-thermal-conductivity substrates. Our results show that considering cooling as an integral part of device design can result in orders-of-magnitude improvements in cooling performance. We use this embedded-cooling approach to demonstrate a super-compact GaN-on-silicon integrated alternating–direct current (a.c.–d.c.) converter, containing four power devices on the same microfluidic-cooled chip, and yielding a power density of 25 kW dm^{-3}. A simple multi-layered printed circuit board (PCB) was designed to direct the coolant flow into the semiconductor device.

Co-design concept and fabrication

Our co-design approach, in which each heat source is coupled to an individual buried cooling-channel serving as a local heat sink, is particularly of interest in GaN power electronic applications with a lateral high-electron-mobility transistor (HEMT) structure. Typical source–drain spacing for HEMTs in >1-kV applications matches the optimum dimensions for microchannel cooling of about 20 µm (refs. [19,42–44]). Therefore, we investigated a GaN-on-Si device structure in which liquid impinges directly onto the epilayer below each contact, ensuring minimum thermal resistance between the hotspot and coolant. In this structure (Fig. 1a), the GaN epilayer provides the power electronics (Fig. 1b), and the silicon functions as a microchannel cooling and fluid-distribution network in a three-dimensional arrangement

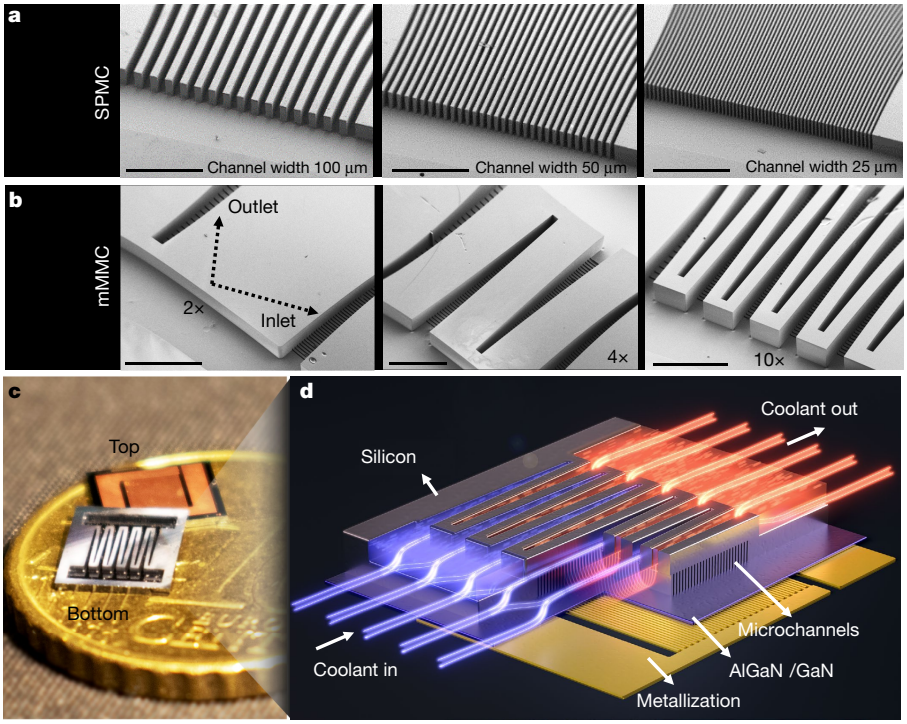


Fig. 2 | Microchannel cooling configurations. a, SEM images of the back side of the silicon substrate with SPMCs. Microchannel widths are 100 µm, 50 µm and 25 µm and the scale bars represent 1 mm. **b**, SEM images of the back side of the silicon substrate with mMMCs, with 2×-, 4×- and 10×-manifold sections. Scale bars represent 850 µm. **c**, Picture of the co-designed devices, from the top and bottom sides, with the 10×-manifold mMMC cooling. The top side shows the electronic structure and the bottom shows the manifold etched in the silicon substrate. **d**, Illustration of the fluid flow through the mMMC structure. Blue lines indicate the cold coolant flow entering the chip, and red lines indicate the hot coolant leaving the chip. The Supplementary Video that visualizes the fluid flow and the three-dimensional render in Fig. 2d were produced by Vytautas Navikas and used with permission.

(Fig. 1c). Figure 1d illustrates the corresponding fabrication method. A staggered pattern of slits was formed in the Si by anisotropic deep etching through narrow incisions in the AlGaN/GaN epilayer to achieve the desired microchannel depth. This pattern provided better structural integrity of the epilayer during fabrication compared to continuous slits. During the subsequent isotropic gas-etch, the channels widened and coalesced in the silicon substrate, while being monitored through the transparent GaN epilayer using an in situ optical etch-rate tracking. This two-step etching process provides independent control over channel width and depth, making it suitable to a wide range of contact pitches. The incisions were finally hermetically sealed during the device metallization step. The Methods and Extended Data Fig. 1 explain the fabrication procedure in detail. Figure 1e shows a scanning electron microscope (SEM) image of the device after the metallization step with sealed channels. Because of the narrow incisions in the epilayer, the contacts do not require substantial oversizing. The microchannels are in direct contact with the active area of the chip, thus providing excellent thermal coupling between the hotspot and the cooling channel (Fig. 1f). Through micrometre-sized openings in the epilayer, 125-µm-deep and 20-µm-wide channels were created in the silicon substrate (Fig. 1g, h).

A series of devices was fabricated with SPMCs with equal width and spacing of 100 µm, 50 µm and 25 µm, and a channel depth of 250 µm in GaN-on-Si power devices, functioning as reference heat sinks (Fig. 2a) for evaluating the performance of the co-designed electronic-microfluidic mMMC devices. Three mMMC chips with 2, 4 and 10 inlet and outlet manifold channels and identical 20 × 125 µm microchannels were fabricated, referred to as the 2×-, 4×- and 10×-manifold chips (Fig. 2b). Figure 2c shows a picture of the mMMC device with the 10×-manifold, including a schematic (Fig. 2d) to illustrate the flow path with coolant impinging directly onto the bottom of the GaN epilayer.

Thermo-hydraulic evaluation

A thermo-hydraulic analysis, using de-ionized water as a coolant, was performed on the cooling structures (Fig. 2a, b) to assess the cooling performance by measuring the thermal resistance, pressure drop and the resulting cooling coefficient of performance (COP), which indicates the energy efficiency of the heat sink. Figure 3a shows the total thermal resistance (R_{total}) between the surface temperature rise and the inlet temperature for the evaluated structures. By reducing the SPMC channel dimensions from 100 µm to 25 µm at identical flow rates, R_{total} reduces, which can be attributed to the increased surface area for heat transfer. However, the 4×- and 10×-manifold heat sinks show an additional substantial reduction in R_{total} compared to the 25-µm SPMC, approaching the limit of single-phase water-cooling (defined by its heat capacity). R_{total} was separated into three components: the contribution due to the heating of the water based on its heat capacity (R_{heat}), the contribution due to convective heat transfer in the microchannels (R_{conv}), and the contribution due to conduction (R_{cond}). The full data reduction procedure to obtain these values is explained in the Methods and in Extended Data Fig. 3. A breakdown of R_{total} is shown in Fig. 3b, revealing a strong relation between R_{conv} and microchannel size, where smaller channels reduce R_{conv}. A large decrease in R_{conv} was achieved with the 10×-manifold, resulting in an 85% and 76% reduction compared to the 50-µm and 100-µm-wide SPMCs, respectively. In combination with a very low R_{cond} for the co-designed manifolds, at a flow rate of 1.0 ml s^{-1}, a thermal resistance of 0.43 K W^{-1} was achieved. The 10×-manifold design thus allows heat fluxes up to 1,723 W cm^{-2} for a maximum temperature rise of 60 K, which is more than twice that of a 25-µm-wide SPMC.

Narrow channels, however, require a higher pressure to achieve equal flow rate (Fig. 3c). For a flow rate of 0.5 ml s^{-1}, SPMC widths of 100 µm, 50 µm and 25 µm require pressures of 160 mbar, 260 mbar

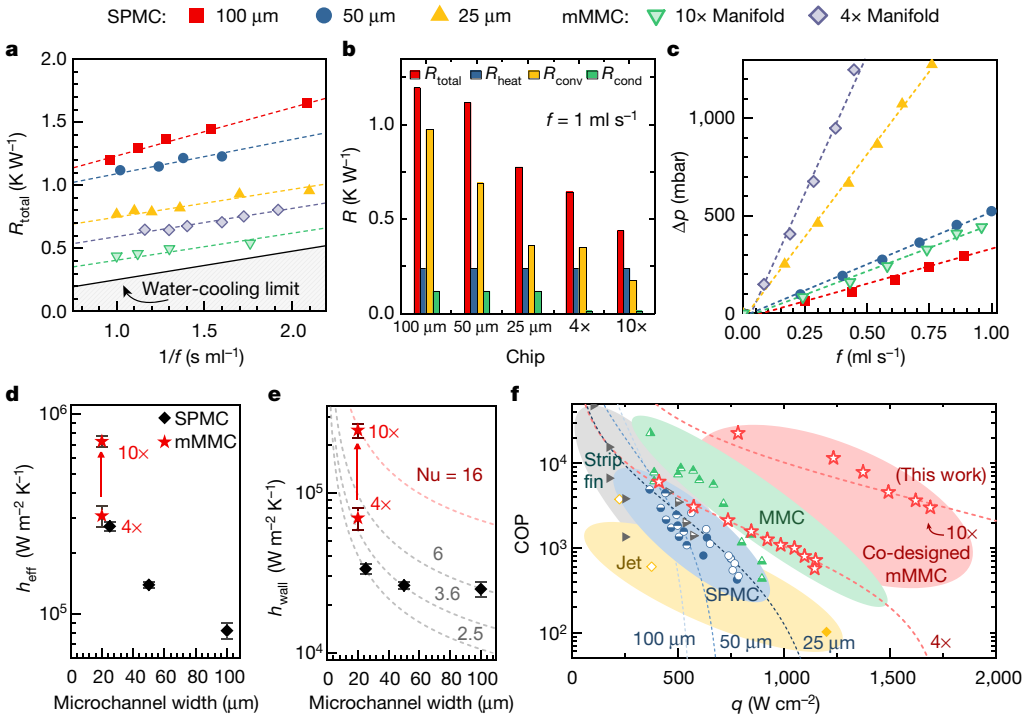


Fig. 3 | Thermo-hydraulic evaluation of the cooling strategies. a, Total thermal resistance R_{total} between the surface temperature and the inlet temperature of the coolant. The black line indicates the lower limit of thermal resistance for single-phase water cooling, determined by its heat capacity. **b**, Breakdown of the contributions of R_{total}, into R_{heat}, R_{conv} and R_{cond}, for all devices evaluated. **c**, Pressure drop Δp versus flow rate f for the considered microchannel structures. The 4×- and 10×-manifold had 20-µm-wide channels. **d**, Effective (base-area averaged) heat-transfer coefficient h_{eff} for straight (SPMC) and manifold (mMMC) structures. Error bars indicate the standard deviation in the heat-transfer coefficient due to averaging of R_{conv} over the measured range of flow rates. **e**, Wall-averaged heat-transfer coefficient h_{wall} for straight and manifold microchannels. A 4.4-fold increase in Nusselt number was observed between the 4×- and 10×-manifold. **f**, Benchmark of the experimentally demonstrated COP versus the maximum heat flux q for a temperature rise of 60 K. Shown are the SPMC (blue), the MMC (green), the impinging jet (yellow), the strip fin (grey) and the mMMC (red). More extensive benchmarking with simulation and analytical results, full references and further classification is provided in Extended Data Fig. 6 and Extended Data Table 2. A large improvement in COP for a given heat flux is achieved with our proposed mMMC structures (red). Dashed lines are models for COP versus heat flux, under the assumption of a constant heat-transfer coefficient and a linear pressure–flowrate relation, fitted through the experimental data.

and 810 mbar, respectively. The manifold structure substantially lowers the pressure drop by reducing the length of the flow path through the microchannel. When splitting the flow into smaller sections with the 10×-manifold, the pressure drop reduced to just 210 mbar. This highlights the benefit of the MMC structure: a lower thermal resistance than SPMCs can be obtained at a reduced pumping power consumption. However, although the manifold structure can reduce the pressure drop, the additional contractions and turns of the fluid can hinder this reduction. For example, 20-µm-wide microchannels in a 4×-manifold require a higher pressure of 1,300 mbar compared to the 25-µm-wide SPMC, which in part can also be attributed to the higher fluid velocity given that the mMMC channels (125 µm) are not as deep as in the SPMC (250 µm). These findings demonstrate the need for a carefully optimized geometry of the microchannel and manifold.

Figure 3d shows a clear trend for SPMCs of increased effective base-area-averaged heat-transfer coefficient (h_{eff}) for smaller microchannels. This is due to the combined effect of the increased surface area and local heat-transfer coefficient in the fully developed laminar-flow regime. The co-designed 4×-mMMC structures matches this trend with $h_{eff} = 3.1 \times 10^5$ W m^{-2} K^{-1}, but a large deviation from this pattern is observed when the effective length through which the coolant flows in the microchannel is reduced. For the 10×-manifold, h_{eff} more than doubles to 7.3×10^5 W m^{-2} K^{-1}, a rise that can be attributed to the high Nusselt number owing to the developing flow in the MMC structure[27,45]. This effect becomes more pronounced when considering the wall-area-averaged heat-transfer coefficient (h_{wall}) (Fig. 3e), which eliminates the contribution of the increased surface area from the heat-transfer coefficient, as well as accounts for the limited fin efficiency of the channels. Over a threefold increase in h_{wall} is observed between 25-µm-wide straight microchannels and the 10×-manifold heat sinks, up to 2.4×10^5 W m^{-2} K^{-1}. This value corresponds to a very high Nusselt number of 16, generally only achieved in larger-scale systems, or in more complex two-phase cooling systems, highlighting the superior thermal performance of this structure.

The combination of improved heat transfer and reduced pressure drop leads to much lower pumping power requirements. The cooling COP is defined as the ratio of extracted power to the pumping power required to provide such a level of cooling, while maintaining a maximum surface temperature rise of 60 K. Higher heat fluxes require higher flow rates, reducing the COP owing to the larger pumping power required. Figure 3f benchmarks the evaluated devices, along with other technologies found in the literature. For SPMC, channel widths of 100 µm, 50 µm and 25 µm show a consecutively higher COP for higher heat fluxes, with a COP in the range between 10^2 and 10^4 and heat fluxes between 350 W cm^{-2} and 800 W cm^{-2}. The 10×-manifold device vastly outperforms these SPMCs. At an identical COP of 5.0×10^3, the 10×-manifold can sustain heat fluxes up to 1.7 kW cm^{-2} at 1.0 ml s^{-1}, compared to 400 W cm^{-2}, 450 W cm^{-2} and 550 W cm^{-2} for the 100-µm, 50-µm and 25-µm SPMCs, respectively. Furthermore, at a heat flux of 780 W cm^{-2}, the 10×-manifold provides a 50-fold increase in COP with respect to 25-µm SPMCs. Compared to MMC heat sinks presented in the literature, the proposed mMMC device outperforms the current state of the art, and demonstrates a large potential for energy-efficient cooling by having a thermal-centred approach in the device design.

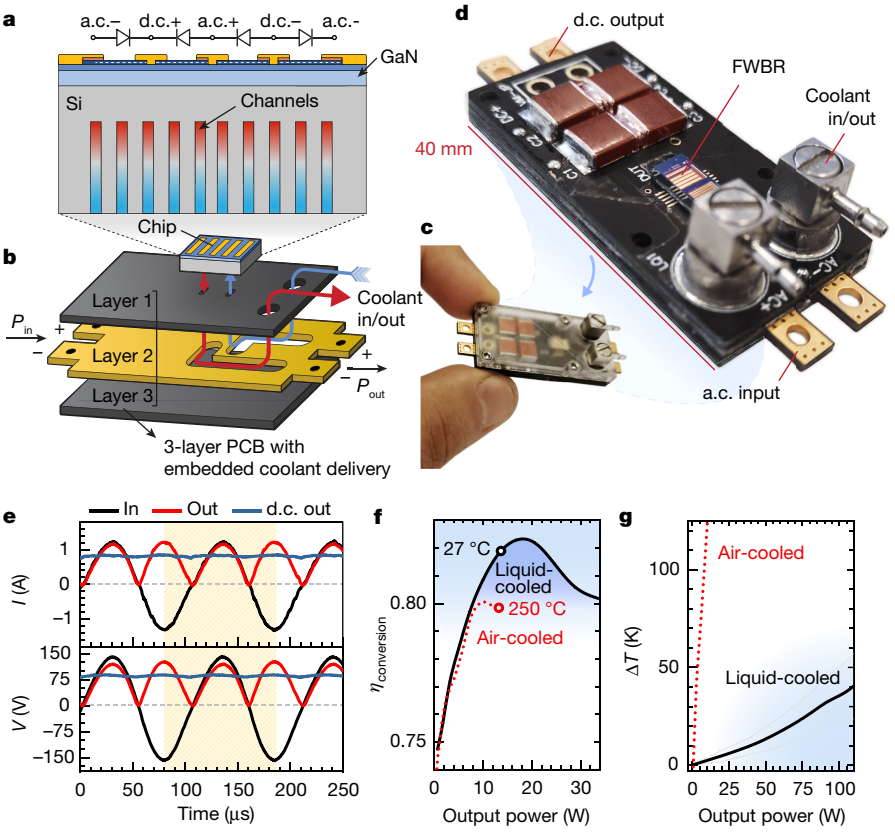


Fig. 4 | a.c.–d.c. converter with embedded liquid cooled GaN power integrated circuit. a, Schematic illustration of the super-compact liquid-cooled power integrated circuit based on four GaN power Schottky barrier diodes integrated in a single chip in a full-bridge configuration. **b**, A PCB-embedded coolant delivery was developed to feed the coolant to the device. The PCB consists of three layers, where the middle layer contains a fluid distribution channel. **c**, Photograph (taken by R.v.E.) of the full 120-W a.c.–d.c. converter with coolant delivery to the liquid-cooled power integrated circuit. **d**, Converter without encapsulation, revealing the monolithically integrated full-wave bridge rectifier (FWBR) integrated circuit. **e**, Rectification waveforms of the converter, 150-V a.c. input (black) and output before (red) and after (blue) filtering using output capacitors. **f**, Efficiency versus output power for the air-cooled and liquid-cooled a.c.–d.c. converter. At identical output power, the liquid-cooled converter exhibits substantially higher efficiency owing to the elimination of self-heating degradation. **g**, Temperature rise versus output power, showing a much higher temperature at equal output power for the air-cooled device compared to the embedded liquid cooling, which causes a large self-heating degradation. The black line shows the mean surface-temperature rise and the highlighted area shows the range between the minumum and maximum temperatures over the device's surface.

Power integrated circuit with embedded cooling

The lateral nature of AlGaN/GaN electronics enables the monolithic integration of multiple power devices onto a single substrate. This opens up opportunities for power electronics, whereby an entire converter can be integrated on a small chip, with large potential for energy, cost and space savings. However, the resulting high heat fluxes limit the maximum output power of the chip. To demonstrate the potential of embedded cooling in a semiconductor device, we monolithically integrated a full-bridge rectifier onto a single GaN-on-Si die. Rectification was provided using four high-performance tri-anode Schottky barrier diodes with a breakdown voltage of 1.2 kV and high-frequency capability up to 5 MHz (ref. [46]). 50-µm-wide cooling channels were integrated on the silicon substrate (Fig. 4a). To fully benefit from the compactness of high-performance microchannel cooling, a three-layer PCB with embedded coolant delivery channels was developed and used to guide the coolant to the device (Fig. 4b). The full fabrication of this monolithically integrated power device and the PCB is described in the Methods and shown in Extended Data Fig. 8. The device was finally fluidically connected to the PCB using laser-cut liquid- and solvent-resistant double-sided adhesive, providing a leak-proof connection. This method is low-cost and easy-to-prototype, and translates well to conventional solder bonding. Figure 4c, d shows the converter implemented, with a very compact form factor, rectifying an a.c. signal with peak voltage and current of 150 V and 1.2 A, respectively (Fig. 4e). Integrated liquid cooling led to a small temperature rise of 0.34 K per watt of output power. For a maximum temperature rise of 60 K, this single die can thus produce an output power of 176 W at a flow rate of only 0.8 ml s^{-1}. Furthermore, the reduced operating temperature led to an increased conversion efficiency (Fig. 4f) by eliminating self-heating degradation from the electrical performance. The a.c.–d.c. converter was experimentally evaluated up to 120 W of output power, while the temperature rise stayed below 50 K (Fig. 4g). Considering the small converter volume (4.8 cm^3), this corresponds to a high power density of 25 kW dm^{-3}. Moreover, since all cooling occurs within its footprint, multiple devices can be densely packed onto the same PCB to increase the output power. This is a clear benefit over conventional heat sinks relying on heat spreading to large areas. These results show that the proposed high-performance cooling approach can enable the realization of high-power (kilowatt range) converters of the size of USB sticks in the foreseeable future.

Discussion and outlook

We present an approach for co-designing microfluidics and electronics for energy-efficient cooling, and demonstrate it on GaN-on-Si power devices by turning the passive silicon substrate from a low-cost carrier into a high-performance heat sink. COP values above 10,000 for heat

fluxes surpassing 1 kW cm^{-2} could be obtained by focusing on cooling in an early stage of the device design. As a practical implication, the average added-energy expenditure of more than 30% for cooling in data centres could potentially drop below 0.01% by adopting this design approach. The entire mMMC cooling structure can be monolithically integrated within the substrate, requiring only conventional fabrication procedures, thus making this economically viable. To realize this concept, solutions for the packaging and interconnects are required. The PCB-based fluid delivery presented provides an example of a way to use these co-designed chips, based on components familiar to the electronics designer. This means that, in order to provide maximum energy savings, cooling should be an integral step in the entire electronic design chain, from the device to the PCB design, and not merely an afterthought. If these practicalities can be addressed, we anticipate that the co-design of microfluidic and electronics will be appropriate for energy-efficient thermally-aware electronics design. This may aid in solving critical challenges in electronics applications, as well as enabling future integrated power converters on a chip to support the electrification of our society in a sustainable manner.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2666-1.

1. Haensch, W. et al. Silicon CMOS devices beyond scaling. *IBM J. Res. Develop*. **50**, 339–361 (2006).
2. Kanduri, A. et al. A perspective on dark silicon. In *The Dark Side of Silicon: Energy Efficient Computing in the Dark Silicon Era* 3–20 (Springer, 2017).
3. Hardavellas, N., Ferdman, M., Falsafi, B. & Ailamaki, A. Toward dark silicon in servers. *IEEE Micro* **31**, 6–15 (2011).
4. Nowak, E. J. Maintaining the benefits of CMOS scaling when scaling bogs down. *IBM J. Res. Develop*. **46**, 169–180 (2002).
5. Garimella, S. V. et al. Thermal challenges in next-generation electronic systems. *IEEE Trans. Compon. Packag. Technol*. **31**, 801–815 (2008).
6. Ohashi, H. Recent power devices trend. *J. Inst. Electr. Eng. Jpn* **122**, 168–171 (2002).
7. *Digitalization And Energy* 103–122 (International Energy Agency, 2017).
8. Shehabi, A. et al. *United States Data Center Energy Usage Report LBNL-1005775* https://www.osti.gov/biblio/1372902 (US DOE, Office of Scientific and Technical Information, 2016).
9. Jones, N. How to stop data centres from gobbling up the world's electricity. *Nature* **561**, 163–166 (2018).
10. Agostini, B. et al. State of the art of high heat flux cooling technologies. *Heat Transf. Eng*. **28**, 258–281 (2007).
11. Chu, R. C., Simons, R. E., Ellsworth, M. J., Schmidt, R. R. & Cozzolino, V. Review of cooling technologies for computer products. In *IEEE Transactions on Device and Materials Reliability* Vol. 4, 568–585 (IEEE, 2004).
12. *2015 Residential Energy Consumption Survey (RECS)* https://www.eia.gov/consumption/residential/data/2015/ (US Energy Information Administration, 2015).
13. DeOreo, W. B., Mayer, P., Dziegielewski, B. & Kiefer, J. *Residential End Uses of Water Version 2* https://www.waterrf.org/resource/residential-end-uses-water-version-2 (Water Research Foundation, 2016).
14. Annual Estimates of the Resident Population for Incorporated Places of 50,000 or More, Ranked by July 1, 2018 Population: April 1, 2010 to July 1, 2018 https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html (United States Census Bureau, 2019).
15. *All-Island Generation Capacity Statement 2018–2027* http://www.soni.ltd.uk/media/documents/Generation_Capacity_Statement_2018.pdf (EirGrid Group, 2018).
16. Amano, H. et al. The 2018 GaN power electronics roadmap. *J. Phys. D* **51**, 163001 (2018).
17. Ohashi, H. et al. Power electronics innovation with next generation advanced power devices. *IEICE Trans. Commun*. **E87-B**, 3422–3429 (2004).
18. Wei, T. et al. High-efficiency polymer-based direct multi-jet impingement cooling solution for high-power devices. *IEEE Trans. Power Electron*. **34**, 6601–6612 (2019).
19. Tuckerman, D. B. & Pease, R. F. W. High-performance heat sinking for VLSI. *IEEE Electron Device Lett*. **2**, 126–129 (1981).
20. Mundinger, D. et al. Demonstration of high-performance silicon microchannel heat exchangers for laser diode array cooling. *Appl. Phys. Lett*. **53**, 1030–1032 (1988).
21. Phillips, R. J. Microchannel heat sinks. *Lincoln Lab. J*. **1**, 31–48 (1988).
22. Harpole, G. M. & Eninger, J. E. Micro-channel heat exchanger optimization. In *1991 Proc. Seventh IEEE Semiconductor Thermal Measurement and Management Symp*. 59–63 (IEEE, 1991).
23. Copeland, D., Behnia, M. & Nakayama, W. Manifold microchannel heat sinks: isothermal analysis. *IEEE Trans. Compon. Packag. Manuf. Technol*. **20**, 96–102 (1997).
24. Copeland, D., Takahira, H., Nakayama, W. & Pak, B. C. Manifold microchannel heat sinks: theory and experiment. In *Advances in Electronic Packaging Proc. Int. Intersociety Electronic Packaging Conference (INTERpack '95)* Vol. 10-2, 829–835, https://library.epfl.ch/beast?record=ebi01_prod001476708 (American Society of Mechanical Engineers, Electrical and Electronics Packaging Division, 1995).
25. Copeland, D. Manifold microchannel heat sinks: numerical analysis. In *Cooling and Thermal Design of Electronic Systems 1995 ASME International Mechanical Engineering Cong. Exp*. Vol. 16, 111–116, https://library.epfl.ch/beast?record=ebi01_prod001583961 (American Society of Mechanical Engineers, Electrical and Electronics Packaging Division, 1995).
26. Copeland, D., Behnia, M. & Nakayama, W. Manifold microchannel heat sinks: conjugate and extended models. *Int. J. Microelectron. Packag. Mater. Technol*. **1**, 139–152 (1998).
27. Mandel, R., Shooshtari, A. & Ohadi, M. A '2.5-D' modeling approach for single-phase flow and heat transfer in manifold microchannels. *Int. J. Heat Mass Transf*. **126**, 317–330 (2018).
28. Ng, E. Y. K. & Poh, S. T. Investigative study of manifold microchannel heat sinks for electronic cooling design. *J. Electron. Manuf*. **9**, 155–166 (1999).
29. Ryu, J. H., Choi, D. H. & Kim, S. J. Three-dimensional numerical optimization of a manifold microchannel heat sink. *Int. J. Heat Mass Transf*. **46**, 1553–1562 (2003).
30. Sarangi, S., Bodla, K. K., Garimella, S. V. & Murthy, J. Y. Manifold microchannel heat sink design using optimization under uncertainty. *Int. J. Heat Mass Transf*. **69**, 92–105 (2014).
31. Cetegen, E., Dessiatoun, S. & Ohadi, M. Heat transfer analysis of force fed evaporation on microgrooved surfaces. In *Proc. 6th Int. Conf. on Nanochannels, Microchannels, and Minichannels (ICNMM2008)* Part A, 657–660, https://asmedigitalcollection.asme.org/ICNMM/proceedings-abstract/ICNMM2008/48345/657/335936 (2008).
32. Kermani, E., Dessiatoun, S., Shooshtari, A. & Ohadi, M. M. Experimental investigation of heat transfer performance of a manifold microchannel heat sink for cooling of concentrated solar cells. In *Proc. Electronic Components and Technology Conf*. 453–459, https://ieeexplore.ieee.org/document/5074053 (IEEE, 2009).
33. Drummond, K. P. et al. A hierarchical manifold microchannel heat sink array for high-heat-flux two-phase cooling of electronics. *Int. J. Heat Mass Transf*. **117**, 319–330 (2018).
34. Back, D. et al. Design, fabrication, and characterization of a compact hierarchical manifold microchannel heat sink array for two-phase cooling. *IEEE Trans. Compon. Packag. Manuf. Technol*. **9**, 1291–1300 (2019).
35. Escher, W., Brunschwiler, T., Michel, B. & Poulikakos, D. Experimental investigation of an ultrathin manifold microchannel heat sink for liquid-cooled chips. *J. Heat Transfer* **132**, 081402 (2010).
36. Schlottig, G. et al. Lid-integral cold-plate topology: integration, performance, and reliability. *J. Electron. Packag*. **138**, 010906 (2016).
37. Robinson, A. J., Kempers, R., Colenbrander, J., Bushnell, N. & Chen, R. A single phase hybrid micro heat sink using impinging micro-jet arrays and microchannels. *Appl. Therm. Eng*. **136**, 408–418 (2018).
38. Everhart, L. et al. Manifold microchannel cooler for direct backside liquid cooling of SiC power devices. In *ASME 5th Int. Conf. on Nanochannels, Microchannels, and Minichannels* 285–292 (ASME, 2007).
39. Gambin, V. et al. Impingement cooled embedded diamond multiphysics co-design. In *2016 15th IEEE Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* 1518–1529 (IEEE, 2016).
40. Drummond, K. P. et al. Evaporative intrachip hotspot cooling with a hierarchical manifold microchannel heat sink array. In *2016 15th IEEE Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* 307–315 (IEEE, 2016).
41. Shekhar Sharma, C. et al. Energy efficient hotspot-targeted embedded liquid cooling of electronics. *Appl. Energy* **138**, 414–422 (2015).
42. Samalam, V. K. Convective heat transfer in microchannels. *J. Electron. Mater*. **18**, 611–617 (1989).
43. Weisberg, A., Bau, H. H. & Zemel, J. N. Analysis of microchannels for integrated cooling. *Int. J. Heat Mass Transf*. **35**, 2465–2474 (1992).
44. Ryu, J. H., Choi, D. H. & Kim, S. J. Numerical optimization of the thermal performance of a microchannel heat sink. *Int. J. Heat Mass Transf*. **45**, 2823–2827 (2002).
45. Shah, R. K. & London, A. L. *Laminar Flow Forced Convection In Ducts : A Source Book For Compact Heat Exchanger Analytical Data* (Academic Press, 1978).
46. Nela, L., Kampitsis, G., Ma, J. & Matioli, E. Fast-switching tri-anode Schottky barrier diodes for monolithically integrated GaN-on-Si power circuits. *IEEE Electron Device Lett*. **41**, 99–102 (2019).

Methods

Device fabrication process

The fabrication process of the co-designed microfluidic-electronic device is shown in Extended Data Fig. 1. Fabrication started with an AlGaN/GaN-on-silicon wafer with, from top to bottom: 2.9-nm-thick GaN cap layer, 20-nm-thick AlGaN barrier, 420-nm-thick GaN channel, 4.2-μm-thick buffer layer, on a 400-μm-thick silicon layer. First, a mesa was etched to define the active area of the chip, followed by a 1-μm-thick plasma-enhanced chemical vapour deposition of SiO_2 as an etching mask to obtain sharp sidewalls after GaN etching. Photoresist was lithographically patterned on top of the SiO_2 layer, to define and open a staggered pattern of slits in the SiO_2 mask with inductively coupled plasma etching using C_4F_8 chemistry. The staggered pattern, with 30-μm-long slits spaced 2 μm apart, prevented the epilayer from turning into a fragile cantilever after performing an undercut in the silicon substrate. Instead, the 2 μm spacing between each slit kept the epilayer together, resulting in good mechanical integrity of the epilayer during the fabrication process. The photoresist was stripped using an O_2 plasma and the exposed GaN slits were consecutively etched using Cl_2+Ar chemistry until the silicon substrate was reached, which was confirmed using end-point detection. The chips were then dipped into 40% KOH at 60 °C for 5 min to remove any remaining AlN-based material from the buffer[47,48]. The Bosch process was used to etch the silicon slits approximately 115 μm deep, resulting in high-aspect-ratio slits. The microchannels in silicon were widened using an isotropic XeF_2 gas etch, which provided selectivity over GaN[49]. XeF_2 gas etching was performed in a pulsed manner: the sample was exposed to XeF_2 at a controlled pressure (1.33 mbar) for 30 s, followed by evacuation of the etching chamber. This process was repeated for 45 cycles until the desired channel width was obtained. In situ optical etching tracking through the transparent GaN membrane was performed using a camera directly mounted on the etching chamber, as shown in Extended Data Fig. 1. This method enabled us to obtain the desired channel width accurately, and to ensure that all slits were coalesced into continuous channels underneath the epilayer. In this way, 20-μm wide microchannels were etched through the narrow openings in the epilayer. Next, the SiO_2 hard mask was stripped using 50% HF for 10 min, and the surface was further cleaned from all organic residues using piranha treatment. A Ti/Al/Ti/Ni/Au Ohmic contact stack was deposited using electron-beam evaporation and photolithographically patterned by lift-off, followed by an annealing step at 850 °C. The inlet and outlet channels were etched into the back side of the chip using the Bosch process, until the channels from both sides coalesced, which was confirmed by optical microscopy. The slits in the GaN epilayer were then sealed by electroplating approximately 7 μm of copper on top of the Ohmic contacts. For the electroplating process, a uniform seed layer of chromium-copper (20 nm/70 nm) was deposited on top of the device after the contact metallization step using electron-beam evaporation, where chromium served as an adhesion layer and copper as the seed layer. Next, 10 μm of photoresist was patterned to define the area to be electroplated. Electrical contact was made with the chip, which functions as the cathode, using an electrically conductive adhesive that was applied over all edges of the chip. First, the chip was briefly dipped in H_2SO_4 to remove any surface oxidation. Then, electroplating was performed using a galvanostat at 1 A for 7 min in a solution containing $CuSO_4$, H_2SO_4 and Cl^-, as well as an addition of Intervia 8510 (Dow), while using a CuP anode. As the galvanically deposited copper film grows conformally and isotropically, the incisions in the GaN layer seal as the copper layer bridges the gap and coalesces on top of the cavity. After electroplating, the photoresist was stripped, and the seed layer was etched by performing a short copper wet-etch ($(NH_4)_2S_2O_8 + H_2SO_4$), followed by a chromium etch that is selective over copper ($KMnO_4 + Na_3PO_4$). Finally, the individual dies were separated using a dicing saw. The Supplementary Video illustrates the flow path of the coolant through this mMMC heat sink structure.

Experimental setup for evaluation of cooling performance

An open-loop single-phase liquid cooling setup, schematically shown in Extended Data Fig. 2a was built underneath an infrared camera in order to perform liquid cooling experiments, as can be seen in Extended Data Fig. 2b. A reservoir of de-ionized water was pressurized with compressed air using a pressure controller (Elveflow OB1 MK3), causing it to flow towards the test section manifold machined out of polyetheretherketone (PEEK) (Extended Data Fig. 2c). PEEK was chosen because of its low thermal conductivity, preventing the heat flux from leaking out of the system by conduction, as well as because of its high glass-transition temperature of 143 °C (ref. [50]). The flow rate of the coolant was measured using a thermal mass flow sensor (Sensirion SLQ-QT500). Chips were mounted on laser-cut poly(methyl methacrylate) (PMMA) carriers with double-sided adhesive and connected to the test section using laser-cut silicone gaskets. A closed seal was obtained on these gaskets using four screws that push down on the PMMA carriers. In this way, no force needs to be applied directly on the chips, preventing the chips from breaking during mounting. Two pressure sensors (Elveflow MPS) were used to measure the pressure at the inlet and outlet of the chip, and the inlet and outlet fluid temperatures are measured using a type-K thermocouple (Thermocaox), integrated just before the inlet and just after the outlet of the chip. The thermocouples were calibrated using a thermostatic bath (Lauda RP855). The chips were connected to a power supply (TTI QPX1200), which simultaneously applies a voltage and measures the current over the device under test (DUT). Electrical connection with the device under test was made using six high-current-rated spring-loaded pins, connected to a custom-made PCB with a hole in the centre to allow infrared measurements. The temperature rise on the surface of the chip was measured using a FLIR SC3000 infrared camera. A LabVIEW automation program was developed to automate the data acquisition. The program waits for the liquid outlet temperature to stabilize, then sends a trigger signal to the video card of the personal computer connected to the infrared camera to record 20 snapshots, and increases the power dissipated on the chip until a critical surface temperature was reached. The surface of the chip was painted black using spray paint to increase emissivity. To further improve the accuracy of the infrared thermography, a pixel-by-pixel emissivity calibration was performed by flowing water at a controlled temperature using the thermostatic bath following the method described in ref. [51]. Infrared emission was measured at each temperature and a fit between temperature and infrared emission was established for each pixel of the photodetector. Finally, a MATLAB script was developed to automate the post-processing of the infrared data, which gave the mean surface temperature rise and the maximum surface temperature rise. The latter was defined as the mean value of the 20 pixels with the highest temperature readings, to be less susceptible to noise.

Pressure test

Before evaluating the cooling performance, pressure tests were performed by increasing the system pressure up to 4 bar (above atmospheric) on each chip. This procedure was intended as a burst test, but no failure was observed up to the maximum pressure capability of the experimental facility. It should be noted that typical epitaxial growth of AlGaN/GaN on a silicon substrate using metal–organic chemical vapour deposition is performed at temperatures around 1,000 °C. Owing to the mismatch in coefficient of thermal expansion, the resulting stress in the epilayer is in typically on the order of 0.3 GPa, whereas the critical cracking stress lies around 1.1 GPa (ref. [52]). Although the additional pressure inside the channels during liquid flow does contribute to the total stress in the epilayer, 1 bar (typical operation) is equivalent to only 0.1 MPa. This stress is more than three orders of magnitude smaller than the typical residual stress in the epilayer, and is therefore not expected

to cause failure. This finding agrees with our observations, as well as with other works in the literature[53,54].

Data reduction

The cooling performance of all chips was analysed for power dissipations up to 75 W and flow rates of 0.1–1.1 ml s^{-1}. Extended Data Fig. 3 shows an overview of the data reduction procedure for the 10×-manifold chip to obtain the relevant values in Fig. 3. The maximum surface temperature ($\Delta T_{surface}$) rise was calculated by subtracting the coolant inlet temperature from the maximum infrared-measured surface temperature (Extended Data Fig. 3a). The liquid temperature rise (ΔT_{liquid}) was calculated by subtracting the inlet water temperature from the water outlet temperature, measured by thermocouple (Extended Data Fig. 3b). The wall temperature (ΔT_{wall}) was calculated by subtracting the mean water temperature between the inlet and outlet from the average surface temperature rise, and performing a correction for one-dimensional conduction through the epilayer, thermal boundary resistance and silicon in case of the straight channels (Extended Data Fig. 3c). A thermal boundary resistance between the GaN and silicon substrate of 1.0×10^{-7} W^{-1} m^2 K was assumed[55–57]. The effective applied power was calculated using an energy balance ($P = f\rho c_p \Delta T_{liquid}$), where ρ and c_p are the density and heat capacity of water, respectively. For all flow rates, the total thermal resistance (R_{total}), the caloric thermal resistance (R_{heat}) and convective thermal resistance (R_{conv}) were determined through a linear fit of the surface temperature rise (Extended Data Fig. 3a), coolant temperature rise (Extended Data Fig. 3b) and wall temperature rise (Extended Data Fig. 3c) versus dissipated power, respectively. Thus, every point in Extended Data Fig. 3d was derived from a wide range of measurements to ensure high accuracy. This figure was plotted against the inverse flow rate to highlight the linear relationship between R_{heat} and f^{-1}. As can be seen, most of the variation of R_{total} with flow rate can be accounted to $R_{heat,}$ whereas R_{conv} shows little dependence on the flow rate. COP was calculated by dividing the maximum heat flux for a ΔT_{max} of 60° temperature rise by the required pumping power (P_{pump}) to achieve this level of cooling[58] ($COP = \Delta T_{max}/P_{pump}R_{total}$), where pumping power was calculated as the product of flow rate and pressure drop ($P_{pump} = f\Delta p$). The effective base-area-averaged heat transfer coefficient was calculated using $h_{eff} = (R_{conv}A_{device})^{-1}$, where A_{die} represents the footprint area of the active area of the device, containing both the electric device and the cooling structure. The average local heat transfer coefficient (h_{wall}) was determined by taking the fin efficiency (η) into account, which was calculated using $\eta = 1$ as a starting point for iteratively solving[59,60]

$$\eta = \frac{\tanh\left(z\sqrt{\frac{2h_{wall}}{k_{Si}w_{wall}}}\right)}{z\sqrt{\frac{2h_{wall}}{k_{Si}w_{wall}}}}.$$

Here, z represents the channel depth, w_{wall} is the channel wall width and k_{Si} is the thermal conductivity of the silicon substrate, which was chosen to be 150 W m^{-1} K^{-1}. Finally, based on h_{wall}, the average Nusselt number (Nu) was calculated for each measurement condition using $Nu = h_{wall}D_h/k_{water}$, where D_h is the hydraulic diameter of the channel ($D_h = (2w_c z)/(w_c + z)$) and k_{water} is the thermal conductivity of water at the mean measured temperature. Extended Data Fig. 5 shows a complete overview of the remaining datasets for temperature rise and thermal resistance of the 25 µm/50 µm/100 µm-SPMC and 4×-manifold MMC, and the full overview of the design parameters and derived values is presented in Extended Data Table 1. Extended Data Fig. 3e shows the Nusselt number and fin efficiency over the measured range of flow speeds, and Extended Data Fig. 3f shows both the effective base-area-averaged and average local heat transfer coefficients. In thermally developing laminar internal flow, the observed average Nusselt number is expected to increase with flow rate, owing to the increased entrance length. At higher flow rates, a longer entrance length will result in a higher heat transfer coefficient. This general trend is observed in Extended Data Fig. 4d. For the 10×-manifold, this effect saturates, probably owing to the short length of the channels, which in combination with a potential shift in coolant distribution over the chip at higher Reynolds number, causes the Nusselt number to peak. A complete overview of the fin efficiencies and Nusselt numbers for all devices can be found in Extended Data Fig. 4c, d. Extended Data Fig. 5 shows the additional thermo-hydraulic analysis on all evaluated devices used for deriving their cooling performance. Extended Data Table 1 summarizes all dimensions and cooling performance of the chips. The performance of the mMMC chips, as well as the SMPC chips evaluated in this work, were benchmarked against a wide range of works in the literature that use water as a coolant (Extended Data Fig. 6). The cooling approaches were classified as SPMC ([19,61]), pin-fins ([61,62]), strip-fins ([61,63,64]), MMC ([29,35,65–68]), impinging jet ([37,61,69–71]), and mMMC (this work). A distinction was made between techniques where the water is in direct contact with the die and the die contains cooling structures (embedded cooling), approaches where the water is in direct with the die, but the die itself does not contain cooling structures (bare-die cooling), and indirect cooling, which requires an additional thermal interface between the heat sink and the chip. All the data used in the benchmarking study are available in the Supplementary Table.

Impact of hydrostatic pressure on electrical performance

Owing to the piezoelectric properties of GaN, changes in pressure and the resulting strain in the epilayer may affect the electrical performance of the device[53,72]. To investigate these phenomena, the outlet of the test section in Extended Data Fig. 2 was plugged. The hydrostatic pressure applied to the test section was swept from 0 mbar to 1,590 mbar and back. At each step in pressure, a cyclic current–voltage measurement was performed, together with the measurement of the water temperature in the test section. After the water reached the ambient temperature of 22 °C, the next measurement was performed. This was done to prevent any drift in temperature during the 3-h-long measurement, which might affect the resistance of the chip. The 14 current–voltage characteristics (Extended Data Fig. 7a) show no clear impact on device performance. Next, R was derived from the current–voltage curves using a linear fit at each pressure condition. The observed variation in R remained within 1.5% of its initial value at atmospheric pressure (Extended Data Fig. 7b). These results show that the effect of the pressure range considered here on the electrical properties of the devices is negligible for the purpose of this work. The small impact of this effect on electrical performance could be attributed to the fact that the microchannels are positioned below the pads, and covered with metal. Any change in carrier density in this region of the chip due to strain would not noticeably affect the device performance, as most of the contribution to the device's resistance occurs in the area between the pads.

a.c.–d.c. converter fabrication

Tri-anode Schottky barrier diode full-wave bridge rectifiers were fabricated on an AlGaN/GaN-on-silicon wafer with, from top to bottom: a 2.9-nm-thick GaN cap layer, a 20-nm-thick AlGaN barrier, a 420-nm-thick GaN channel and a 4.2-µm-thick buffer layer on a 400-µm-thick silicon substrate. The tri-anode/tri-gate regions were first defined using electron-beam lithography with a width and spacing of 200 nm, followed by a 200-nm-deep inductively coupled plasma etch following the process previously described in ref. [46]. These dimensions have been shown to result in high breakdown voltage and excellent on-state performance[73]. After Ohmic metal deposition for the cathode contacts, 20-nm-thick SiO_2 was deposited by atomic layer deposition as the tri-gate dielectric, and then selectively removed in the tri-anode region. A Ni/Au metal stack was deposited onto the tri-gate/tri-anode region to form the Schottky contact, as well as on the cathode. Extended Data Fig. 8d shows a SEM image of four scaled-up tri-gate Schottky barrier diodes forming the full-wave bridge rectifier. The close-up SEM image

shows the Schottky barrier diode structure. The channel length was 16.5 µm, corresponding to 1.2 kV of breakdown voltage[46]. Next, the wafer was temporarily bonded to a carrier wafer before microchannels were etched in the back side using deep reactive ion etching to a depth of approximately 500 µm. After detaching the substrate from the carrier wafer and dicing, the individual liquid-cooled full-wave bridge rectifier was attached to a three-layer PCB using water-resistant adhesive with embedded coolant delivery channels. The top layer of the PCB provides the electric circuit connections and the middle layer contains the coolant delivery channels (Extended Data Fig. 8a). The individual layers of the PCB were easily connected using laser-cut adhesive (Extended Data Fig. 8b). Pressure-sensitive double-sided adhesive was used from AR-Global (ARseal 90880), with water- and solvent-resistant properties as well as a high-temperature operation range (up to 120 °C). A rectangular piece with inlet and outlet holes was laser-cut using a CO_2 laser. The double-sided adhesive was placed on the PCB, aligning the inlet and outlet holes of the adhesive with the PCB. Next, the chip was attached to adhesive on the PCB to create a seal. This approach emphasizes the ability to assemble a prototype without the need of expensive machines. Alternatively, since the PCB contains a gold-plated metalized landing pad, conventional large-scale industrial processes can be used as well, such as (eutectic) solder bonding between a metallization layer on the backside of the chip and the PCB. Extended Data Fig. 8c shows the final assembled converter.

a.c.–d.c. converter evaluation

The cooling performance of the a.c.–d.c. converter was investigated by connecting all four Schottky barrier diodes in parallel, such that a uniform known d.c. power dissipation could be applied to the chip. For flow rates varying between 0.08 ml s^{-1} and 0.8 ml s^{-1}, the surface temperature rise was monitored increasing power dissipation up to 25 W (Extended Data Fig. 9b). The flow-rate-dependent thermal resistance was derived from the slope of surface temperature versus power (Extended Data Fig. 9c). For each flow rate, the pressure drop between the inlet and outlet was measured, and the corresponding pumping power was calculated (Extended Data Fig. 9d). Over the entire range of measured flow rates, the total pumping power stayed below 62 mW, which can be easily supplied by miniaturized piezoelectric micropumps to achieve a high system-level power density. To study the power-conversion performance of the a.c.–d.c. converter, the device was connected to a full-bridge inverter with *LC* filter to supply a 100-kHz a.c. input, up to 200 V peak to peak. The d.c. output of the converter was connected to a load of 50 Ω, and the flow rate was fixed at 0.8 ml s^{-1}. Extended Data Fig. 9a shows the input a.c. and output d.c. waveforms of the converter at 70 W of transferred power. Surface temperature was monitored using an infrared camera, while power was increased until a critical surface temperature rise of 60 K was observed. Following this approach, up to 120 W of output power could be delivered using this compact power converter.

Data availability

All the data needed to evaluate the conclusions in the paper are present in the paper, in the Extended Data and in the Supplementary Information.

47. Mileham, J. R. et al. Wet chemical etching of AlN. *Appl. Phys. Lett.* **67**, 1119 (1995).
48. Guo, W. et al. KOH based selective wet chemical etching of AlN, $Al_xGa_{1-x}N$, and GaN crystals: a way towards substrate removal in deep ultraviolet-light emitting diode. *Appl. Phys. Lett.* **106**, 082110 (2015).
49. Vicknesh, S., Tripathy, S., Lin, V. K. X., Wang, L. S. & Chua, S. J. Fabrication of deeply undercut GaN-based microdisk structures on silicon platforms. *Appl. Phys. Lett.* **90**, 071906 (2007).
50. van der Vegt, A. K. & Govaert, L. E. *Polymeren: Van Keten Tot Kunstof* (VSSD, 2003).
51. Szczukiewicz, S., Borhani, N. & Thome, J. R. Fine-resolution two-phase flow heat transfer coefficient measurements of refrigerants in multi-microchannel evaporators. *Int. J. Heat Mass Transf.* **67**, 913–929 (2013).
52. Raghavan, S. & Redwing, J. M. Growth stresses and cracking in GaN films on (111) Si grown by metal–organic chemical-vapor deposition. I. AlN buffer layers. *J. Appl. Phys.* **98**, 023514 (2005).
53. Chapin, C. A., Miller, R. A., Dowling, K. M., Chen, R. & Senesky, D. G. InAlN/GaN high electron mobility micro-pressure sensors for high-temperature environments. *Sens. Actuat. A* **263**, 216–223 (2017).
54. Tan, X. et al. High performance AlGaN/GaN pressure sensor with a Wheatstone bridge circuit. *Microelectron. Eng.* **219**, 111143 (2020).
55. Sarua, A. et al. Thermal boundary resistance between GaN and substrate in AlGaN/GaN electronic devices. *IEEE Trans. Electron Dev.* **54**, 3152–3158 (2007).
56. Turin, V. O. & Balandin, A. A. Performance degradation of GaN field-effect transistors due to thermal boundary resistance at GaN/substrate interface. *Electron. Lett.* **40**, 81–83 (2004).
57. Kuzmík, J. et al. Transient thermal characterization of AlGaN/GaN HEMTs grown on silicon. *IEEE Trans. Electron Dev.* **52**, 1698–1705 (2005).
58. Steinke, M. E. & Kandlikar, S. G. Single-phase liquid heat transfer in plain and enhanced microchannels. In *Proc. 4th Int. Conf. on Nanochannels, Microchannels and Minichannels (ICNMM2006)* https://asmedigitalcollection.asme.org/ICNMM/proceedings-abstract/ICNMM2006/47608/943/323023 (ASME, 2006).
59. Tosun, I. *Modeling in Transport Phenomena* (Elsevier, 2007).
60. Hesselgreaves, J. E., Law, R. & Reay, D. A. *Compact Heat Exchangers* 2nd edn Vol. 1, Ch. 7, 275–360 (Butterworth-Heinemann, 2016).
61. Ndao, S., Peles, Y. & Jensen, M. K. Multi-objective thermal design optimization and comparative analysis of electronics cooling technologies. *Int. J. Heat Mass Transf.* **52**, 4317–4326 (2009).
62. Brunschwiler, T. et al. Interlayer cooling potential in vertically integrated packages. In *Microsystem Technologies* Vol. 15, 57–74 (Springer, 2009).
63. Kandlikar, S. G. & Upadhye, H. R. Extending the heat flux limit with enhanced microchannels in direct single phase cooling of computer chips. In *21st Ann. IEEE Symp. on Semiconductor Thermal Measurement and Management* 8–15 (IEEE, 2005).
64. Colgan, E. G. et al. A practical implementation of silicon microchannel coolers for high power chips. In *21st Ann. IEEE Symp. on Semiconductor Thermal Measurement and Management* 1–7 (IEEE, 2005).
65. Jung, K. W. et al. Embedded cooling with 3D manifold for vehicle power electronics application: single-phase thermal-fluid performance. *Int. J. Heat Mass Transf.* **130**, 1108–1119 (2019).
66. Ohadi, M., Choo, K., Dessiatoun, S. & Cetegen, E. *Next Generation Microchannel Heat Exchangers* 1–111 (Springer, 2013).
67. Han, Y., Lau, B. L., Tang, G., Zhang, X. & Rhee, D. M. W. Si-based hybrid microcooler with multiple drainage microtrenches for high heat flux cooling. *IEEE Trans. Compon. Packag. Manuf. Technol.* **7**, 50–57 (2017).
68. Han, Y., Lau, B. L., Zhang, X., Leong, Y. C. & Choo, K. F. Thermal management of hotspots with a microjet-based hybrid heat sink for GaN-on-Si devices. *IEEE Trans. Compon. Packag. Manuf. Technol.* **4**, 1441–1450 (2014).
69. Ditri, J., Hahn, J., Cadotte, R., McNulty, M. & Luppa, D. Embedded cooling of high heat flux electronics utilizing distributed microfluidic impingement jets. In *ASME 2015 Int. Tech. Conf. Exhib. on Packaging and Integration of Electronic and Photonic Microsystems(InterPACK 2015)/ASME 2015 13th Int. Conf. on Nanochannels, Microchannels, and Minichannels* Vol. 3, T10A014, https://library.epfl.ch/beast?record=ebi01_prod010609314 (ASME, 2015).
70. Natarajan, G. & Bezama, R. J. Microjet cooler with distributed returns. *Heat Transf. Eng.* **28**, 779–787 (2007).
71. Wei, T. et al. High efficiency direct liquid jet impingement cooling of high power devices using a 3D-shaped polymer cooler. In *Technical Digest International Electron Devices Meeting (IEDM)* 32.5.1–32.5.4 https://ieeexplore.ieee.org/document/8268487 (IEEE, 2018).
72. Dzuba, J. et al. AlGaN/GaN diaphragm-based pressure sensor with direct high performance piezoelectric transduction mechanism. *Appl. Phys. Lett.* **107**, 122102 (2015).
73. Ma, J., Santoruvo, G., Tandon, P. & Matioli, E. Enhanced electrical performance and heat dissipation in AlGaN/GaN Schottky barrier diodes using hybrid tri-anode structure. *IEEE Trans. Electron Devices* **63**, 3614–3619 (2016).

Acknowledgements We thank the staff at the Center of Micro and Nano Technology (CMi) for support and advice on the fabrication processes. We thank V. Navikas for his graphical assistance to the paper. This work was supported in part by the European Research Council (ERC Starting Grant) under the European Union's H2020 program (ERC grant agreement number 679425), in part by the Swiss Office of Energy (grant number SI501568-01) and in part by the Swiss National Science Foundation (Assistant Professor Energy grant number PYAPP2_166901).

Author contributions R.v.E. and E.M. conceived the project. R.v.E., R.S. and L.N. developed and optimized device fabrication processes. R.v.E. and L.N. fabricated the devices. R.v.E. designed and developed the experimental setup to study cooling performance. G.K. designed the circuits for evaluating the fabricated devices. R.v.E. and G.K. designed and performed the experiments. R.v.E. analysed the data. E.M. supervised the project. R.v.E. and E.M. wrote the manuscript with input from all authors.

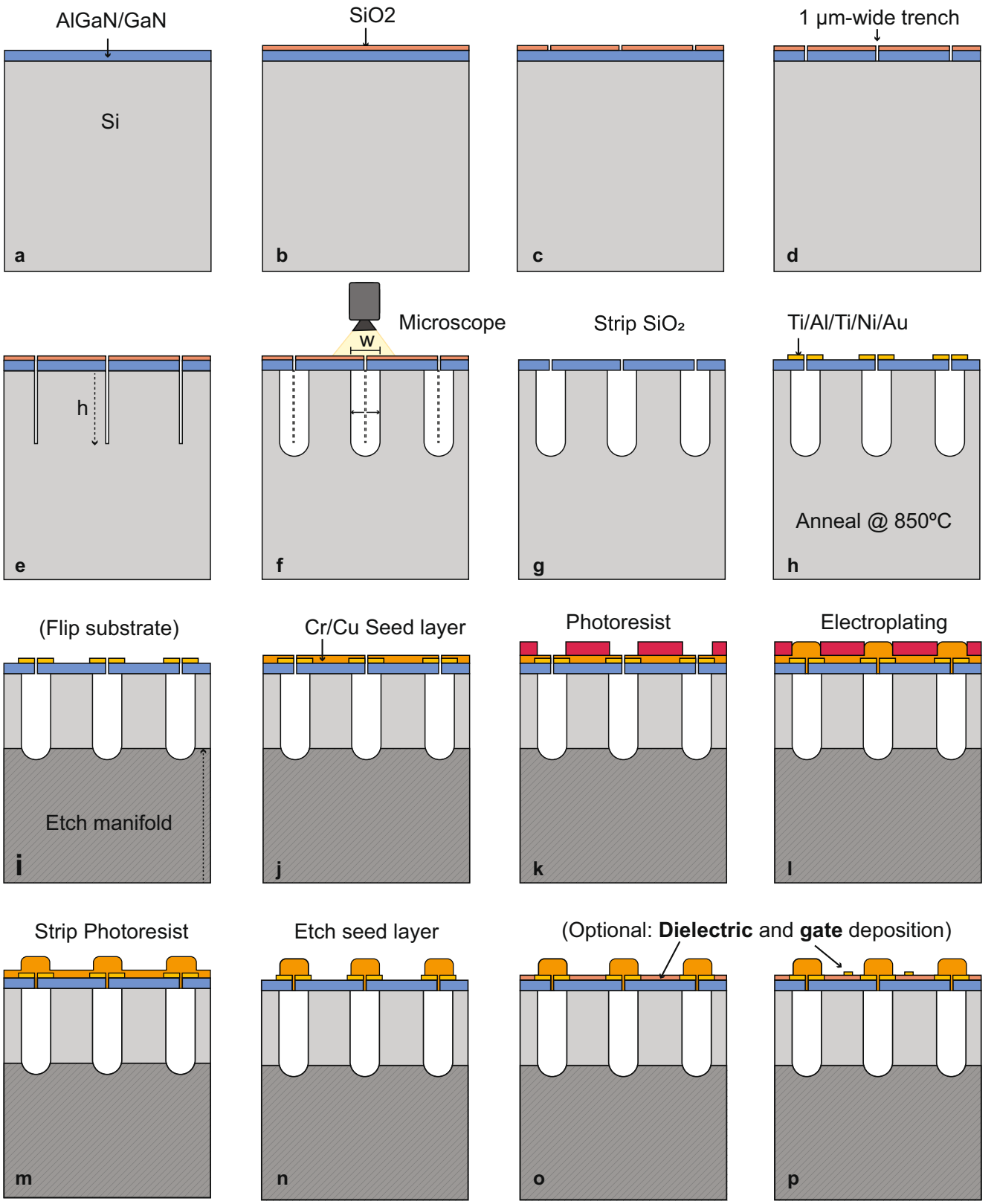
Competing interests The authors declare no competing interests.

Additional information

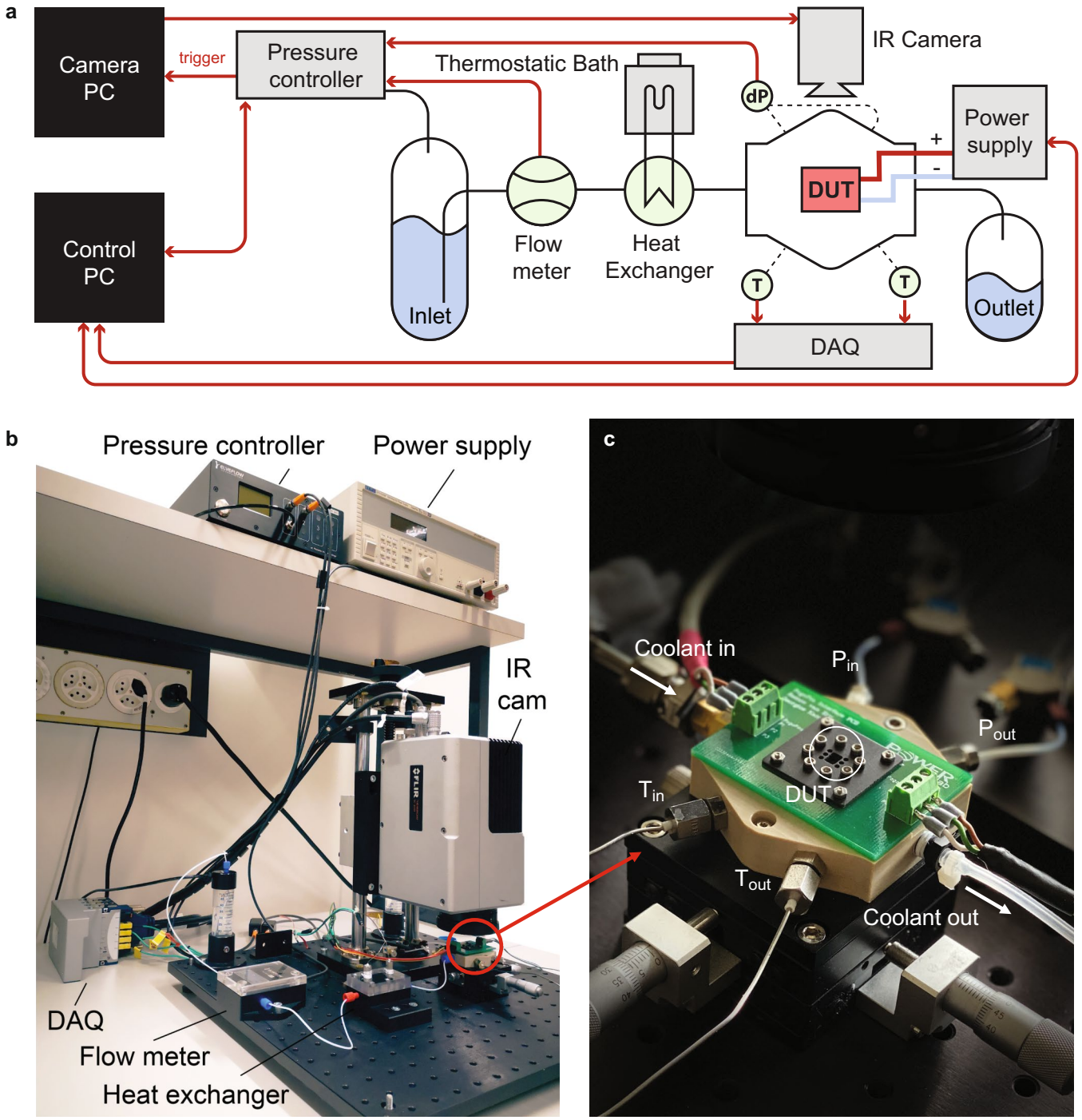
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2666-1.

Correspondence and requests for materials should be addressed to E.M.

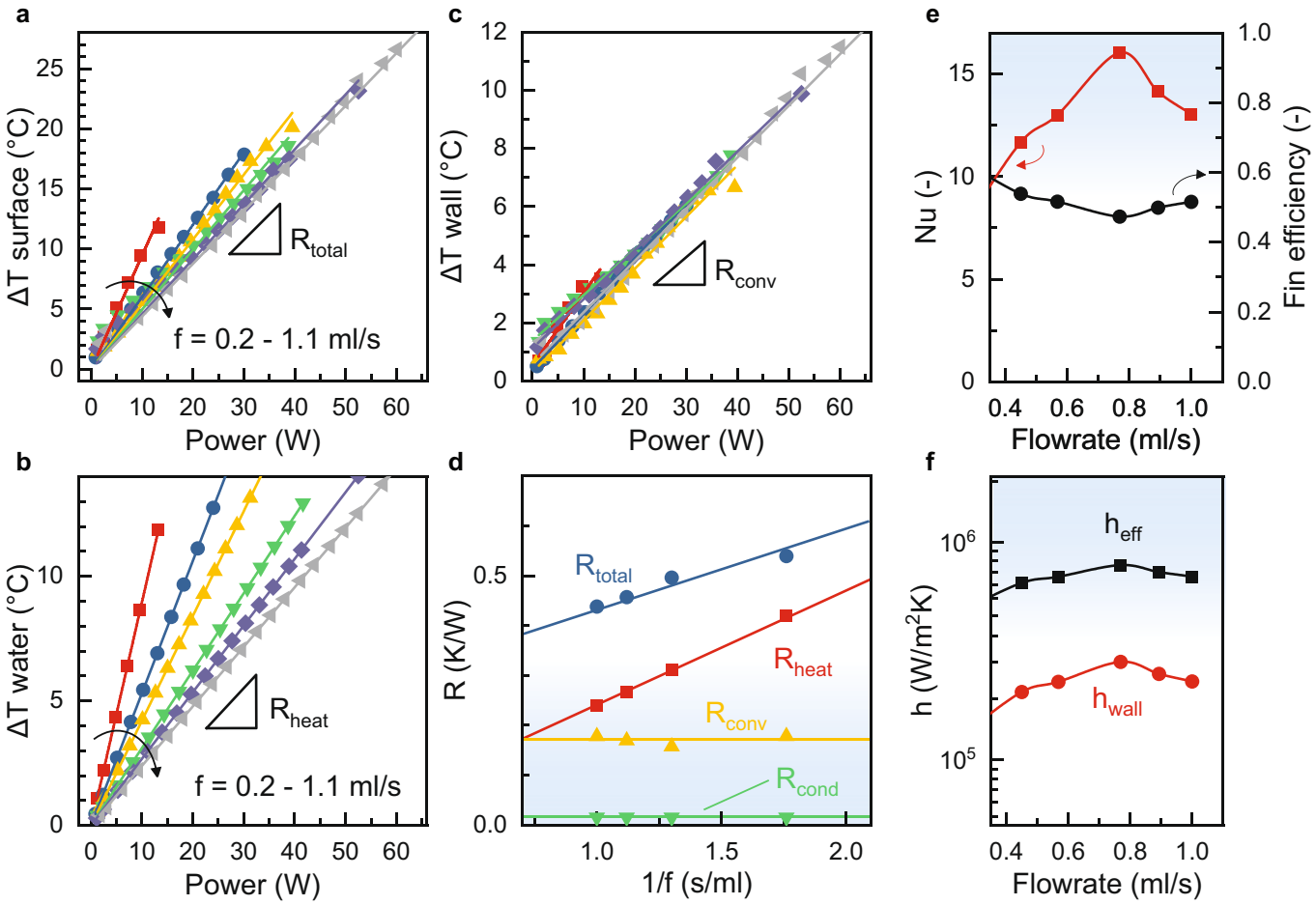
Peer review information *Nature* thanks Peter Moens, Tiwei Wei and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



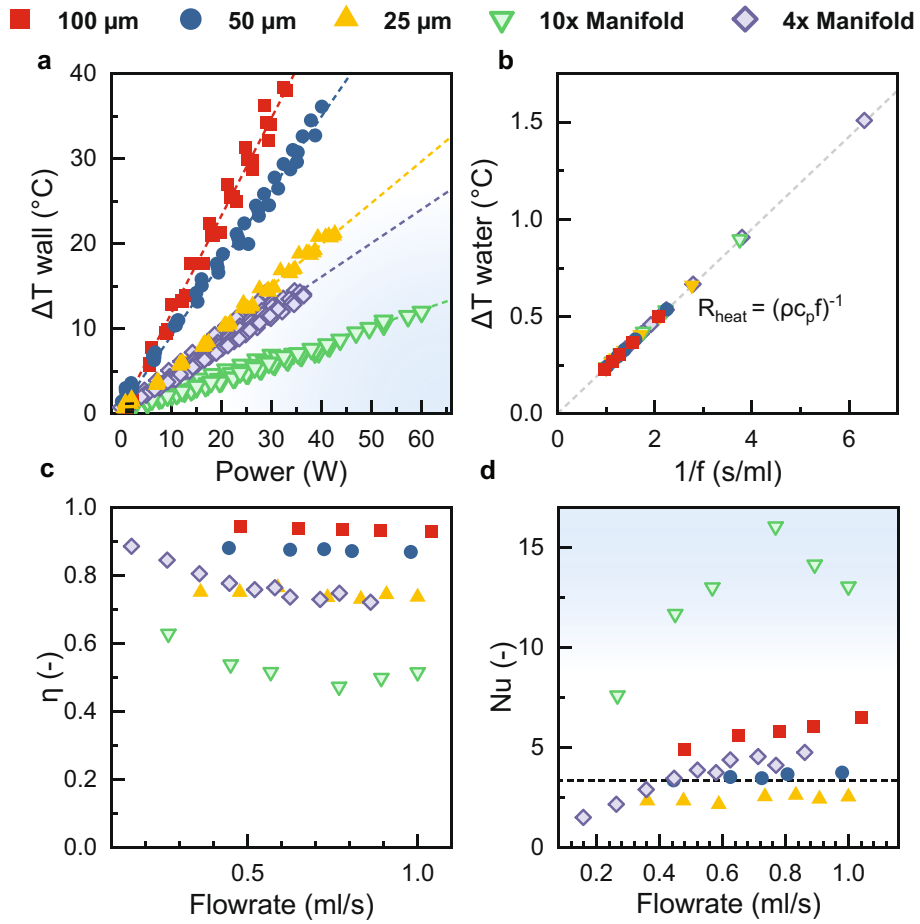
Extended Data Fig. 1 | Fabrication process of the co-designed microfluidic-electric device. a, AlGaN/GaN epilayer on a silicon substrate. **b**, SiO_2 hard-mask deposition. **c**, Hard mask patterning and opening. **d**, Epilayer etching until the substrate is reached. **e**, Anisotropic deep (to depth *h*) etching of the silicon substrate through the epilayer opening. **f**, Isotropic gas etching through the epilayer opening to widen the slits under the epilayer. An in situ optical etching tracking was put in place to control the width *w* of the channels. **g**, Hard-mask removal. **h**, Ohmic contact deposition and annealing, and seed layer deposition for electroplating and patterning the electroplating mask. **i**, Manifold channel etching from the back of the substrate. **j**, Cr/Cu seed layer deposition for electroplating. **k**, Lithography step to define electroplating openings. **l**, Electroplating to seal the epilayer openings. **m**, Photoresist removal. **n**, Wet etch to remove Cr/Cu seed layer. **o**, Finish device fabrication with optional dielectric deposition. **p**, Finish device fabrication with optional gate metal deposition.



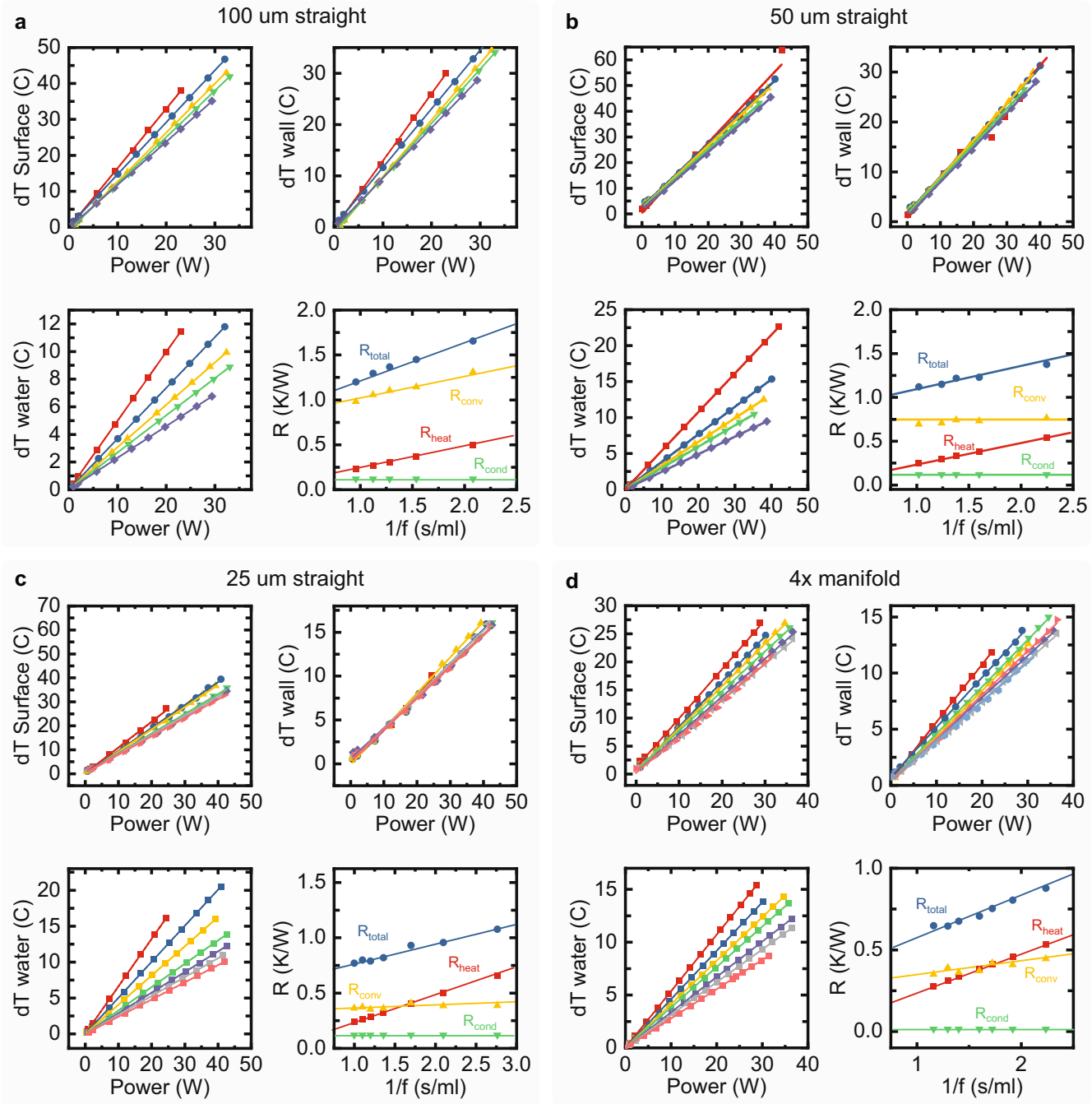
Extended Data Fig. 2 | Experimental setup for evaluating the thermo-hydraulic performance. a, Schematic overview of the measurement setup. An inlet reservoir of coolant is pressurized using a pressure controller, whereas the temperature is controlled using a thermostatic bath. Liquid flow through a flow meter into the test section, containing the chip (DUT). The temperature of the chip is monitored using an infrared (IR) camera, and coolant temperature is monitored using thermocouples (T) and transferred to the personal computer (PC) using a data acquisition box (DAQ). Pressure drop over the chip (dP) is measured at the inlet and outlet port of the chip. **b**, Picture of the experimental setup for characterizing the thermal performance. **c**, Close-up picture of the test section.



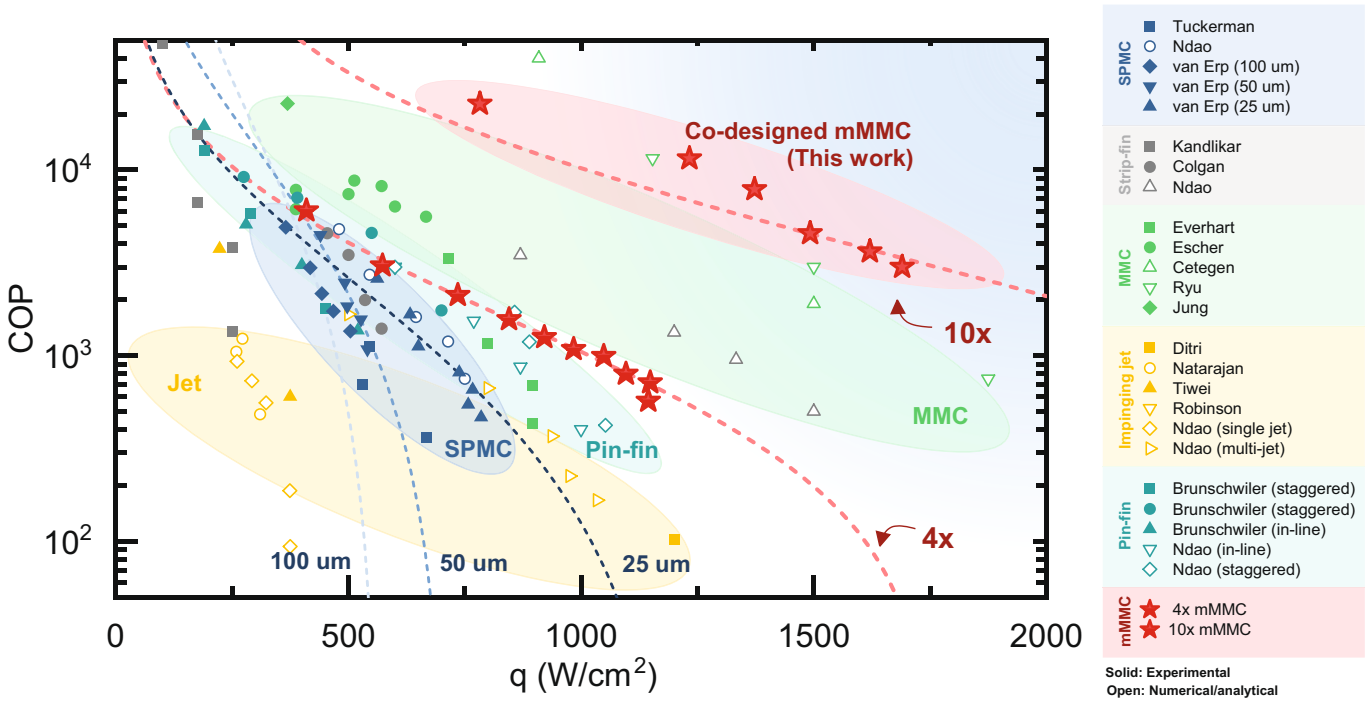
Extended Data Fig. 3 | Example data reduction of thermal characterization experiments for the 10×-manifold chip. a, Peak surface temperature rise above the inlet temperature, measured using infrared thermography at varying power dissipation. The slope of the linear fit through the data points gives the total thermal resistance (R_{total}). **b**, Water temperature rise, measured between the inlet and outlet of the chip. The slope of the linear fit through the data points gives the contribution of the total thermal resistance due to the temperature rise of the water (R_{heat}). **c**, Wall temperature rise. The slope through these data points gives the convective thermal resistance. **d**, Total, caloric, convective and conductive thermal resistance versus the inverse flow rate. **e**, Nusselt number and fin efficiency. **f**, Effective base-area averaged heat transfer coefficient (h_{eff}) and wall-area averaged heat transfer coefficient (h_{wall}), taking the surface area of the microchannels as well as the fin efficiency into account.



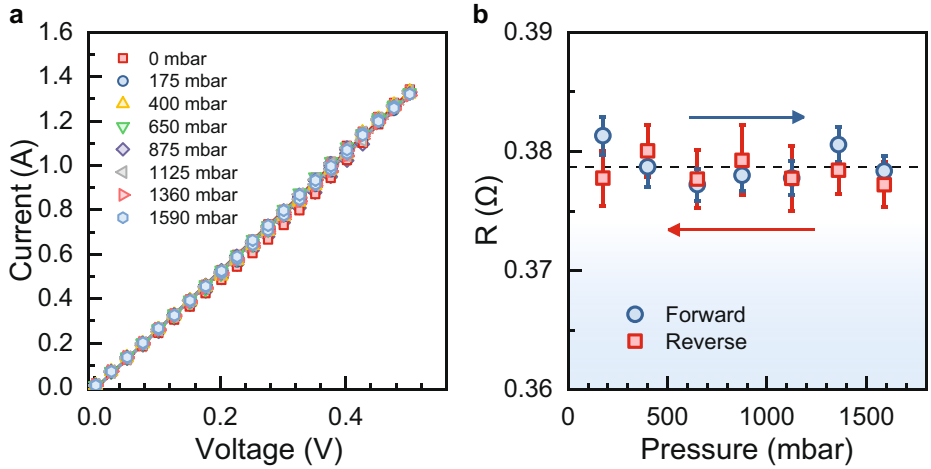
Extended Data Fig. 4 | Overview of derived values of the thermo-hydraulic analysis. a, Wall temperature for all devices. Each device shows a distinct slope in wall temperature rise versus power dissipation. **b**, Caloric thermal resistance for all evaluated flow rates, showing a clear $(\rho c_p f)^{-1}$ relationship over all devices. **c**, Fin efficiency over a range flow rates. **d**, Nusselt number versus inverse flow rate. The dashed line indicates that Nu = 3.66 for fully developed internal flow.



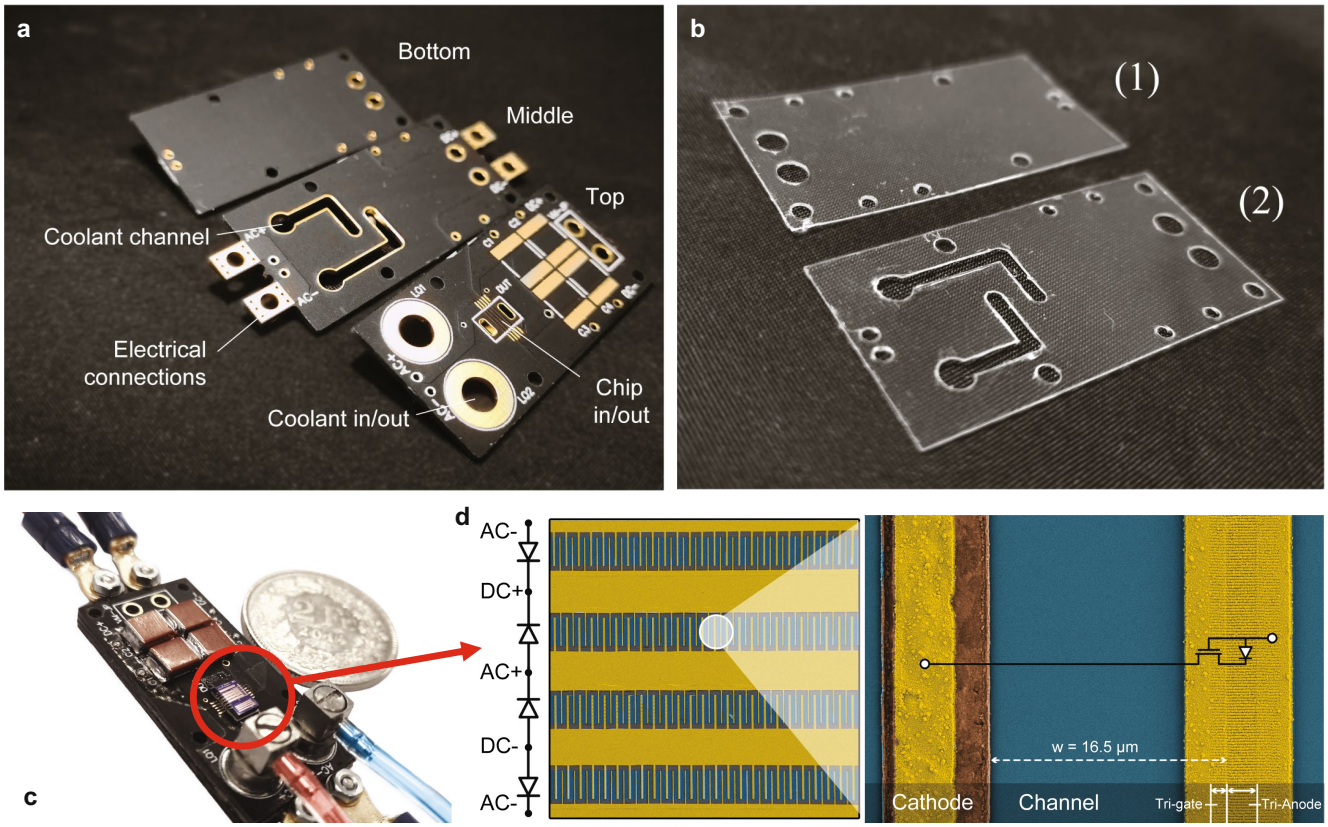
Extended Data Fig. 5 | Additional thermo-hydraulic data. Surface temperature rise dT, wall temperature rise, water temperature rise and thermal resistance R for: **a**, 100 µm-wide SPMC; **b**, 50 µm-wide SPMC straight microchannels; **c**, 25 µm-wide SPMC straight microchannels; and **d**, 4×- manifold.



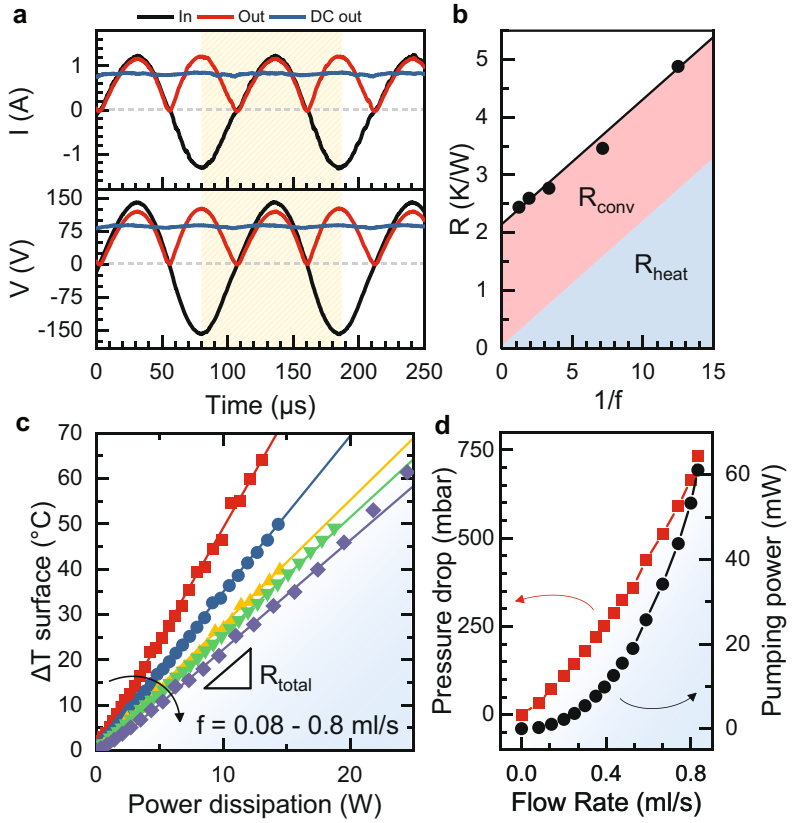
Extended Data Fig. 6 | Extensive benchmarking plot of micro-structured cooling approaches in the literature using water as a working fluid. COP versus heat flux for a maximum surface temperature rise of 60 K. Solid markers indicate experimental results and open markers indicate numerical or analytical calculations. The results in this work are indicated by red stars. Dashed lines correspond to predictions based on a constant heat transfer. References [19,29,35,38,61–66,68–71] to all datasets used can be found in Extended Data Table 2, and all the data used are available in the Supplementary Table.



Extended Data Fig. 7 | Impact of pressure on electrical performance. **a**, Current–voltage characteristics at hydrostatic pressure between 0 mbar and 1,590 mbar. **b**, Normalized change in electrical resistance versus pressure during a sweep in pressure up to 1,600 mbar and back. Each resistance value was extracted using a linear fit ($R^2 > 0.997$) through a cyclic current–voltage measurement from 0 V to 0.5 V and back. Error bars indicate standard error of the fit over each set of 46 data points per condition.

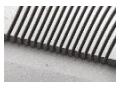
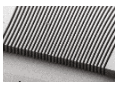
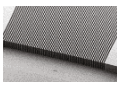


Extended Data Fig. 8 | Structure of the integrated full-bridge rectifier with embedded cooling. a, Three PCBs that provide coolant delivery to the chip. **b**, Laser-cut adhesives used to bond the layers together. **c**, Converter after assembly, with electrical and fluidic connections. **d**, SEM image of the four-diode structure. The inset shows the polarity of each device and a close-up of the structure of the tri-anode Schottky barrier diode.



Extended Data Fig. 9 | Operation of the a.c.–d.c. converter with embedded cooling. a, Input and output waveforms of 150 V/1.2 A peak-to-peak rectification at 100 kHz. **b**, Surface temperature rise versus power dissipation for varying flow-rates. **c**, Thermal resistance versus inverse flow rate of the full converter. Pressure drop and pumping power versus flow rate. **d**, Pressure drop and pumping power versus flow rate.

Extended Data Table 1 | Table of all design parameters and measured values per chip

			SPMC			mMMC	
Parameter		**Unit**					
Manifold channels	N_m	*[-]*	-	-	-	4	10
Channel Width	w_c	*[µm]*	100	50	25	20	20
Channel Depth	z	*[µm]*	250	250	250	125	125
Hydraulic diameter	D_h	*[µm]*	142	83	45	34	34
Device/cooling area	A_d	*[cm^2]*	0.099	0.099	0.099	0.081	0.081
Wetted area	A_{wet}	*[cm^2]*	0.348	0.598	1.09	0.476	0.476
Average convective thermal resistance	R_{conv}	*[K/W]*	1.1	0.73	0.37	0.41	0.17
	R'_{conv}	*[cm^2K/W]*	1.1×10^{-1}	7.2×10^{-2}	3.7×10^{-2}	3.3×10^{-2}	1.4×10^{-2}
Conductive thermal resistance	R_{cond}	*[K/W]*	1.2×10^{-2}	1.2×10^{-2}	1.2×10^{-2}	1.5×10^{-3}	1.5×10^{-3}
	R'_{cond}	*[cm^2K/W]*	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-4}	1.2×10^{-4}
Average effective heat transfer coefficient	h_{eff}	*[W/m^2K]*	8.2×10^4	1.4×10^5	2.7×10^5	3.1×10^5	7.3×10^5
Average local heat transfer coefficient	h_{wall}	*[W/m^2K]*	2.5×10^4	2.6×10^4	3.3×10^4	6.9×10^4	2.4×10^5
Average fin efficiency	η	*[-]*	0.94	0.87	0.74	0.76	0.50
Maximum Nusselt number	Nu	*[-]*	6.5	3.7	2.6	4.7	16
Hydraulic resistance	r_h	*[mbar-s/ml]*	358	535	1692	2191	479

R is thermal resistance in K W^{-1}, and *R'* is surface-area-normalized thermal resistance in cm^2 K W^{-1}.

Extended Data Table 2 | Selected references for the benchmarking study in Extended Data Fig. 6

Ref.	Authors	Year	Approach*	Geometry	Application	Work
This work	van Erp	2020	Embedded	mMMC	Power electronics	Experimental
[19]	Tuckerman & Pease	1981	Embedded	SPMC	Logic	Experimental
[61]	Ndao et al.	2009	-	SPMC	General	Optimization
[61]	Ndao et al.	2009	-	In-line pin fin	General	Optimization
[61]	Ndao et al.	2009	-	Staggered pin fin	General	Optimization
[61]	Ndao et al.	2009	-	Offset strip fin	General	Optimization
[61]	Ndao et al.	2009	-	Single Jet	General	Optimization
[61]	Ndao et al.	2009	-	Multi-Jet	General	Optimization
[38]	Everhart et al.	2007	Bare die	MMC bare die cooling	Power diode	Experimental
[35]	Escher et al.	2010	Embedded	MMC	Logic	Experimental
[64]	Colgan et al.	2005	Embedded or indirect	Offset strip fin	Logic	Experimental
[69]	Ditri et al.	2015	Bare die	Multi-jet	Bare die cooling	Numerical & experimental
[68]	Han et al.	2014	Indirect (Chip included)	Hybrid Jet/MMC	RF power amplifier	Numerical
[63]	Kandlikar & Upadye	2005	-	Offset strip fin	General	Experimental
[70]	Natarajan & Bezama	2007	Direct	Multi-jet	Logic	Numerical
[71]	Wei et al.	2018	Bare die	Multi-jet	Logic	Experiment
[62]	Brunschwiler et al.	2009	Embedded	Pin-Fin Staggered	3D integration	Experimental
[62]	Brunschwiler et al.	2009	Embedded	Distorted Pin-Fin staggered	3D integration	Experimental
[62]	Brunschwiler et al.	2009	Embedded	In-line pin fin	3D integration	Experimental
[29]	Ryu et al.	2003	-	MMC	General	Numerical
[66]	Ohadi et al.	2013	Indirect (Chip not included)	MMC	General	Numerical
[65]	Jung et al.	2019	Direct	MMC	Vehicle power electronics	Experimental

Data are from refs. [19,29,35,38,61–66,68–71]. *'Embedded' means that the liquid is in direct contact with the die, and the cooling structure is embedded inside the die. 'Bare die' means that the liquid is in direct contact with the die, but no cooling structures are embedded inside the die. 'Indirect' means that the liquid is not in direct contact with the die, and no cooling structures are embedded inside the die. An additional interface between the chip and the cooling device is required. RF, radio frequency.

Article

Evidence for supercritical behaviour of high-pressure liquid hydrogen

https://doi.org/10.1038/s41586-020-2677-y

Received: 28 June 2019

Accepted: 10 July 2020

Published online: 9 September 2020

Check for updates

Bingqing Cheng[1,2,3]✉, Guglielmo Mazzola[4], Chris J. Pickard[5,6] & Michele Ceriotti[7,8]

Hydrogen, the simplest and most abundant element in the Universe, develops a remarkably complex behaviour upon compression[1]. Since Wigner predicted the dissociation and metallization of solid hydrogen at megabar pressures almost a century ago[2], several efforts have been made to explain the many unusual properties of dense hydrogen, including a rich and poorly understood solid polymorphism[1,3–5], an anomalous melting line[6] and the possible transition to a superconducting state[7]. Experiments at such extreme conditions are challenging and often lead to hard-to-interpret and controversial observations, whereas theoretical investigations are constrained by the huge computational cost of sufficiently accurate quantum mechanical calculations. Here we present a theoretical study of the phase diagram of dense hydrogen that uses machine learning to 'learn' potential-energy surfaces and interatomic forces from reference calculations and then predict them at low computational cost, overcoming length- and timescale limitations. We reproduce both the re-entrant melting behaviour and the polymorphism of the solid phase. Simulations using our machine-learning-based potentials provide evidence for a continuous molecular-to-atomic transition in the liquid, with no first-order transition observed above the melting line. This suggests a smooth transition between insulating and metallic layers in giant gas planets, and reconciles existing discrepancies between experiments as a manifestation of supercritical behaviour.

Liquid hydrogen constitutes the interior of giant planets and brown dwarf stars, and it is commonly assumed to undergo a first-order phase transition between an insulating molecular fluid and a conducting metallic fluid[1]. Understanding the nature of this liquid–liquid transition (LLT) is crucial for accurately modelling the structure and evolution of giant planets, including Jupiter, Saturn and many exoplanets[8]. Standard planetary models assume a sharp LLT that is accompanied by a discontinuity in density, and therefore give a clear-cut transition between an inner metallic mantle and an outer insulating mantle[9].

Probing the nature of the LLT in the laboratory faces the challenges of creating controllable high-pressure and -temperature environments and of confining hydrogen specimens while making measurements. Consequently, experimental studies have not yet reached a consensus on whether the LLT is a first-order or a smooth transition[10]. Furthermore, there are considerable discrepancies of up to 100 GPa (see Fig. 1a) between experiments at the transition pressure of the LLT[10–15] in the phase diagram.

Given the experimental difficulties, computer simulations have played a fundamental role in characterizing the phase diagram of hydrogen[6,16–18], using a quantum mechanical treatment of electrons to describe atomic interactions. Different levels of electronic-structure theories have been employed, ranging from the accurate quantum Monte Carlo (QMC) methods[17,19,20], to density functional theory (DFT) approximations[16–18,21,22]. Whereas early simulations gave contradictory results[16,19,21], the most recent calculations identify small density discontinuities below 1,500 K (refs. [17,18,20,22]), which have been interpreted as the signatures of a first-order LLT.

Even for DFT simulations, which offer a balance between computational cost and efficiency, the systems studied are limited to sizes of few hundreds of atoms and timescales of a few picoseconds[16–18,21,22]. Given the subtlety of phase-transition phenomena, it is important to overcome the size and timescale limitations, as well as to elucidate the effect of the details of the electronic-structure methods on the location of the LLT, the melting line and the stabilities of the different solid phases[17,18,23].

To address these issues, we constructed three sets of machine-learning potentials (MLPs), using the Behler–Parrinello artificial neural network architecture[24]. The three MLPs are based on different electronic-structure references: DFT with the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional, DFT with the Becke88–Lee–Yang–Parr (BLYP) functional, and variational QMC[20]. More details and benchmarks are provided in Supplementary Information. The results from the three methods are in qualitative agreement. In what follows we report the results generated by the MLP based on PBE, and describe the other results in Supplementary Information. For both solid and liquid structures of small sizes, the MLP shows excellent agreement with the underlying ab initio method. Moreover, the low

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. [2]TCM Group, Cavendish Laboratory, University of Cambridge, Cambridge, UK. [3]Trinity College, Cambridge, UK. [4]IBM Quantum, IBM Research – Zurich, Rüschlikon, Switzerland. [5]Department of Materials Science and Metallurgy, University of Cambridge, Cambridge, UK. [6]Advanced Institute for Materials Research, Tohoku University, Sendai, Japan. [7]Laboratory of Computational Science and Modeling, Institute of Materials, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [8]National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ✉e-mail: bc509@cam.ac.uk

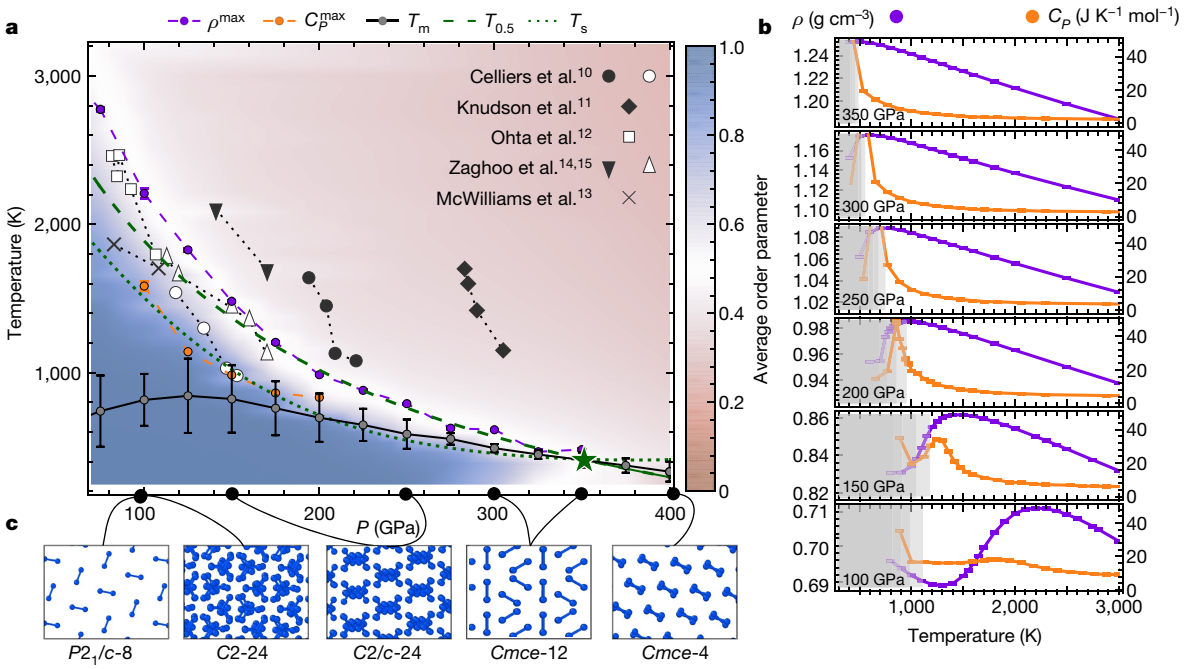


Fig. 1 | Thermodynamic properties of high-pressure hydrogen predicted by the MLP based on PBE DFT. The results from MLPs based on BLYP DFT and QMC are shown in Supplementary Information. **a**, The colour scheme indicates the molecular fraction defined by the order parameter. The black curve is the estimated solid–liquid coexistence line, with the upper and the lower bound of hysteresis indicated by the error bars. The density (ρ) and molar heat capacity (C_P) maxima at different pressures are indicated by purple and orange dots, respectively. The dashed and dotted green curves are the coexistence and phase-separation lines of atomic and molecular fluids predicted by the polyamorphic solution model, respectively. The intersection between the two green curves, marked by a green star, is the predicted location of the critical point of the LLT. The experimental results are taken from refs. [10–15]. **b**, The purple curves show the density isobar, and the orange curves show the molar heat capacity at different pressures. The shaded regions indicate the conditions under which solid phases are stable, corresponding to the solid–liquid coexistence line shown in **a**. Error bars indicate statistical uncertainties. **c**, At each given pressure (black lines), the crystalline structure, the space group and the size of the primitive cell of the solid hydrogen phase with the lowest enthalpy are shown.

cost of the MLP allowed us to investigate hydrogen phase transitions using large system sizes and long simulation times. If performed using DFT, the total computational cost of this study would have required several hundred millions of CPU years, exceeding the capacity of the world's fastest supercomputers.

Solid–liquid transition

Solid hydrogen exhibits complex polymorphism, and only a few of its crystal structures and phase boundaries have been characterized conclusively[1,3–5,25,26]. Encouragingly, the MLP correctly captures the ground-state crystals (Fig. 1c) during random searches[27] over a wide pressure range between 100 GPa and 400 GPa (see details in Supplementary Information). Furthermore, it reproduces the subtle enthalpy differences of a few millielectronvolts between the competing polymorphs[28,29], consistently with the underlying DFT reference. The accuracy of these predictions demonstrates the immense promise of MLPs for crystal structure discovery.

To estimate the melting curve without prior knowledge of the solid-phase diagram, we applied a simple hysteresis method: 1,728-atom hydrogen systems were first cooled from the liquid phase until solidification, and subsequently reheated until melting, for a total of 0.8 ns of molecular dynamics (MD) simulation time. At each pressure we performed eight simulations. Owing to the presence of nucleation barriers, the freezing and melting temperatures are affected by hysteresis. Therefore, the true melting point T_m of the system lies between these two temperatures and can be estimated using their mean value. The shape of the estimated melting line (the black curve in Fig. 1a), which has a peak at around 125 GPa and 850 K and then declines at higher pressures, is in excellent agreement with recent experimental measurements[30] and previous PBE results[6].

Liquid–liquid transition

We performed MD simulations across a broad range of temperatures and pressures using 1,728-atom simulation cells and 0.4 ns simulation time for each run. We employed an order parameter defined as the fraction of atoms that have one neighbour within a smooth cutoff function that is equal to 1 up to 0.8 Å and decays to 0 at 1.1 Å. As evident from Fig. 1a, the molecular fraction varies smoothly across the liquid phase diagram, with the transition region becoming narrower at low temperature (T) and high pressure (P). Other observables, including the density (ρ), the molar heat capacity C_P (Fig. 1c), the pair correlation function and the electronic density of states (see Supplementary Information), also show the absence of discontinuities. Both ρ and C_P exhibit anomalous behaviours, namely, smooth peaks that become sharper at higher pressures (Fig. 1b). The loci of these maxima, as well as the atomic–molecular transition region, converge towards the melting line at above 350 GPa.

Polyamorphic solution model

The MD simulation results (Fig. 1a) suggest the lack of a sharp LLT. To provide a more quantitative analysis, we determined the parameters of a polyamorphic solution model[22,31] that describes a mixture of two interconvertible liquid states. At each thermodynamic state point, the regular-solution molar free energy $g(x)$ as a function of the molecular fraction x is

$$g(x) = x\Delta g + k_B T x \ln x + k_B T(1-x)\ln(1-x) + \omega x(1-x). \quad (1)$$

The term $\Delta g = g_m - g_a$ is the chemical potential difference between the atomic and the molecular phases, ω is an enthalpic term that accounts

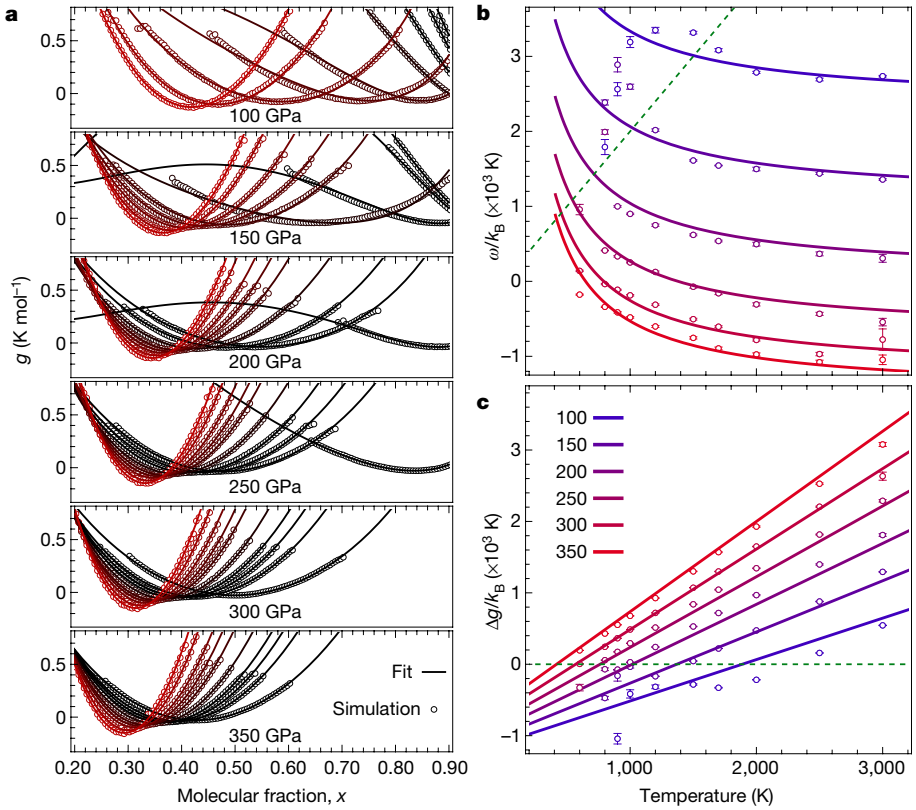


Fig. 2 | Polyamorphic solution model fits of the high-pressure hydrogen system. a, The dots show the computed Gibbs free energy profiles $g(x)$ as functions of the molecular fraction order parameter. The results are from one of the eight sets of metadynamics simulations. The smooth curves show the individual fits to the polyamorphic solution model. The series correspond to results obtained at T = 600, 800, 900, 1,000, 1,200, 1,500, 1,700, 2,000, 2,500 and 3,000 K, plotted in shades of red, from dark to bright. As expected, the minimum of $g(x)$ shifts to lower molecular fractions as the temperature increases. **b**, The dots represent values of $\Delta g = g_m - g_a$ fitted to the solution model, and the lines are linear fits to Δg. **c**, The dots are the individual values of ω obtained from fitting $g(x)$ to the solution model at different pressures and temperatures, and the curves are fits to those values. The dotted green line corresponds to $\omega = 2T$, which corresponds to the phase-separation line, and the dashed line to $\Delta g = 0$, that is, the coexistence line. The error bars were estimated from the error of the mean for the eight sets of simulations.

for the non-ideality of mixing and k_B is the Boltzmann constant. To obtain a free-energy profile that can be compared to equation (1) from simulations, we performed a set of metadynamics[32] simulations in which we enhanced the spontaneous fluctuations of the order parameter x (see Supplementary Information).

As shown in Fig. 2a, each $g(x)$ obtained from simulations has a single minimum and matches perfectly the form predicted by the solution model. This indicates perfect mixing of the two liquids and absence of an LLT throughout the range of temperatures and pressures that we explored. In addition, we used the simple empirical expressions $\Delta g = a_0 + a_1P + a_2T + a_3PT$ and $\omega = b_0 + b_1P + b_2/T + b_3P^2$ to describe the model parameters Δg and ω. As shown in Fig. 2b, c, these expressions agree well with the Δg and ω obtained by independent fits at each state point, at temperature and pressure conditions above the melting line. The small discrepancy at low temperatures and low pressures is caused by solidification, which the solution model does not consider. We then used the analytic expressions to estimate the $x = 0.5$ coexistence line $T_{0.5}$ (that is, the temperature at which atomic and molecular fluids become equally stable, determined by $\Delta g(P, T) = 0$)) and the phase-separation line T_s (that is, the temperature below which the two fluids start demixing, determined as $T_s = \omega(P, T_s)/2k_B$). We note that being at $T < T_s$ is not sufficient to observe the demixing behaviour; it is also necessary that the atomic and molecular phases are equally stable, because the two phases are interconvertible. It is also worth noting that, although there are different ways to define the molecular fraction x using an order parameter, the curves T_s and $T_{0.5}$ are rather insensitive to such definitions. These two curves are plotted in Fig. 1a as dashed and dotted green lines, respectively. The two lines cross at the critical point (marked by a green star) of the fluid–fluid phase transition, which is located at $(P_c, T_c) \approx (350 \pm 1\,\text{GPa}, 416 \pm 2\,\text{K})$, coinciding approximately with the melting line. At $T > T_c$ the system exhibits supercritical behaviour, without phase separation and with anomalies in the thermodynamic properties of the mixture following different Widom lines that emanate from the critical point. At $T < T_c$, the system shows a first-order phase transition, and the $T_{0.5}$ coexistence line becomes the phase boundary. Because for this system T_c approximately coincides with T_m, no sharp LLT can be observed. Instead, the anomalous behaviours induced by this hidden critical point can be observed throughout the liquid phase diagram, much like the case of water[33].

Our observation of the supercritical hydrogen fluid above the melting line contradicts several recent DFT and QMC simulations[17,18,20], which showed a sharp LLT, suggested by small discontinuities in density up to around 1,000–1,500 K. The probable origin of this discrepancy, which we traced to finite-size effects on the solid–liquid transition, is discussed in detail in Supplementary Information. We performed explicit DFT MD simulations with a system size of 128 atoms. We reproduced the pressure–density relations of the previous studies[17,18], albeit using a coarser density grid along each isotherm. However, we observed the formation of solid phases at temperatures up to T = 1,250 K in constant-pressure simulations and the appearance of defective solids up to T = 1,000 K in constant-volume simulations. Such solidification-like events caused discontinuities in the density and radial-distribution functions, similar to a sharp LLT.

The results above are based on the reference PBE potential energy surface, but our conclusion concerning supercriticality is insensitive to the electronic-structure method used to compute the atomic interactions. The analogous results based on the MLPs trained on BLYP DFT and variational QMC datasets are shown in Supplementary Information. The different methods used to obtain the electronic structure lead to different locations of the coexistence and phase-separation lines, but the qualitative picture remains the same. In all cases, the critical point is below the melting line. Our conclusion is reinforced by the analysis based on the polyamorphic solution model. Although a change in the electronic structure method, the inclusion of nuclear quantum effects, and the residual difference between the MLP and the reference method can influence the details of the phase diagram, the solution model shows that in the region that is usually proposed for the critical point, the system is far from critical. At $P < 200$ GPa, $T_{0.5}$ is above T_s, which means that the atomic fluid is unstable at the conditions at which phase separation can happen. At about $P > 200$ GPa, ω is negative at $T \gtrsim 800$ K, suggesting that mixing is enthalpically—and not only entropically—favourable.

The predicted supercritical behaviour of fluid hydrogen can explain the discrepancies between different experiments. If the LLT is indeed first-order, all observables should undergo an abrupt change when crossing the coexistence line. Instead, the supercriticality of fluid hydrogen means that the boundary of the LLT is blurred and its location depends on the specific criterion used to define it. In other words, different observables may exhibit anomalous behaviours that follow different Widom lines, as we observed for the density and heat capacity. Indeed, the LLT boundaries measured by different teams at $1{,}000 < T < 2{,}000$ K all qualitatively extrapolate towards the proposed critical point[10–15] (see Fig. 1a). The observation of a sharper transition in the low-temperature compression experiments of Knudson et al.[11], in comparison to those performed by Celliers et al.[10], is also consistent with supercritical behaviour.

The polyamorphic solution model, which we validated in our simulations by combining an MLP trained on electronic-structure calculations and thorough statistical sampling, models the stability and miscibility of atomic and molecular hydrogen and quantitatively describes the molecular-to-atomic transition in dense liquid hydrogen. This model thus provides a thermodynamic picture of the LLT, and can be directly employed to interpret experiments and astrophysical observations. Our general approach can be used to quantitatively assess the properties of mixtures of hydrogen and heavier elements, as well as to address the long-standing questions concerning Jupiter's core solubility and the anomalous-luminosity problem of Saturn[8]. The accuracy of the MLP for the solid phases demonstrates enormous promise to answer the many open questions concerning the solid-phase diagram of dense hydrogen[26].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2677-y.

1. McMahon, J. M., Morales, M. A., Pierleoni, C. & Ceperley, D. M. The properties of hydrogen and helium under extreme conditions. *Rev. Mod. Phys.* **84**, 1607–1653 (2012).
2. Wigner, E. & Huntington, H. B. On the possibility of a metallic modification of hydrogen. *J. Chem. Phys.* **3**, 764–770 (1935).
3. Howie, R. T., Guillaume, C. L., Scheler, T., Goncharov, A. F. & Gregoryanz, E. Mixed molecular and atomic phase of dense hydrogen. *Phys. Rev. Lett.* **108**, 125501 (2012).
4. Zha, C., Liu, Z., Ahart, M., Boehler, R. & Hemley, R. J. High-pressure measurements of hydrogen phase IV using synchrotron infrared spectroscopy. *Phys. Rev. Lett.* **110**, 217402 (2013).
5. Dalladay-Simpson, P., Howie, R. T. & Gregoryanz, E. Evidence for a new phase of dense hydrogen above 325 gigapascals. *Nature* **529**, 63–67 (2016).
6. Bonev, S. A., Schwegler, E., Ogitsu, T. & Galli, G. A quantum fluid of metallic hydrogen suggested by first-principles calculations. *Nature* **431**, 669–672 (2004).
7. Ashcroft, N. W. Metallic hydrogen: a high-temperature superconductor? *Phys. Rev. Lett.* **21**, 1748–1749 (1968).
8. Guillot, T. The interiors of giant planets: models and outstanding questions. *Annu. Rev. Earth Planet. Sci.* **33**, 493–530 (2005).
9. Hubbard, W. B. & Militzer, B. A preliminary Jupiter model. *Astrophys. J.* **820**, 80 (2016).
10. Celliers, P. M. et al. Insulator-metal transition in dense fluid deuterium. *Science* **361**, 677–682 (2018).
11. Knudson, M. D. et al. Direct observation of an abrupt insulator-to-metal transition in dense liquid deuterium. *Science* **348**, 1455–1460 (2015).
12. Ohta, K. et al. Phase boundary of hot dense fluid hydrogen. *Sci. Rep.* **5**, 16560 (2015).
13. McWilliams, R. S., Dalton, D. A., Mahmood, M. F. & Goncharov, A. F. Optical properties of fluid hydrogen at the transition to a conducting state. *Phys. Rev. Lett.* **116**, 255501 (2016).
14. Zaghoo, M., Salamat, A. & Silvera, I. F. Evidence of a first-order phase transition to metallic hydrogen. *Phys. Rev. B* **93**, 155128 (2016).
15. Zaghoo, M. & Silvera, I. F. Conductivity and dissociation in liquid metallic hydrogen and implications for planetary interiors. *Proc. Natl Acad. Sci. USA* **114**, 11873–11877 (2017).
16. Scandolo, S. Liquid–liquid phase transition in compressed hydrogen from first-principles simulations. *Proc. Natl Acad. Sci. USA* **100**, 3051–3053 (2003).
17. Morales, M. A., Pierleoni, C., Schwegler, E. & Ceperley, D. M. Evidence for a first-order liquid–liquid transition in high-pressure hydrogen from ab initio simulations. *Proc. Natl Acad. Sci. USA* **107**, 12799–12803 (2010).
18. Lorenzen, W., Holst, B. & Redmer, R. First-order liquid–liquid phase transition in dense hydrogen. *Phys. Rev. B* **82**, 195107 (2010).
19. Delaney, K. T., Pierleoni, C. & Ceperley, D. M. Quantum Monte Carlo simulation of the high-pressure molecular-atomic crossover in fluid hydrogen. *Phys. Rev. Lett.* **97**, 235702 (2006).
20. Mazzola, G., Helled, R. & Sorella, S. Phase diagram of hydrogen and a hydrogen–helium mixture at planetary conditions by quantum Monte Carlo simulations. *Phys. Rev. Lett.* **120**, 025701 (2018).
21. Vorberger, J., Tamblyn, I., Militzer, B. & Bonev, S. A. Hydrogen–helium mixtures in the interiors of giant planets. *Phys. Rev. B* **75**, (2007).
22. Geng, H. Y., Wu, Q., Marqués, M. & Ackland, G. J. Thermodynamic anomalies and three distinct liquid–liquid transitions in warm dense liquid hydrogen. *Phys. Rev. B* **100**, 134109 (2019).
23. Clay, R. C. III et al. Benchmarking exchange-correlation functionals for hydrogen at high pressures using quantum Monte Carlo. *Phys. Rev. B* **89**, 184106 (2014).
24. Behler, J. & Parrinello, M. Generalized neural network representation of high-dimensional potential energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
25. Magdău, I. B., Marqués, M., Borgulya, B. & Ackland, G. J. Simple thermodynamic model for the hydrogen phase diagram. *Phys. Rev. B* **95**, 094107 (2017).
26. Loubeyre, P., Occelli, F. & Dumas, P. Synchrotron infrared spectroscopic evidence of the probable transition to metal hydrogen. *Nature* **577**, 631–635 (2020).
27. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
28. Pickard, C. J. & Needs, R. J. Structure of phase III of solid hydrogen. *Nat. Phys.* **3**, **473** (2007).
29. Monserrat, B., Needs, R. J., Gregoryanz, E. & Pickard, C. J. Hexagonal structure of phase III of solid hydrogen. *Phys. Rev. B* **94**, 134101 (2016).
30. Zha, C., Liu, H., Tse John, S. & Hemley, R. J. Melting and high *P–T* transitions of hydrogen up to 300 GPa. *Phys. Rev. Lett.* **119**, 075302 (2017).
31. Anisimov, M. A. et al. Thermodynamics of fluid polyamorphism. *Phys. Rev. X* **8**, 011004 (2018).
32. Laio, A. & Parrinello, M. Escaping free energy minima. *Proc. Natl Acad. Sci. USA* **99**, 12562–12566 (2002).
33. Soper, A. K. & Ricci, M. A. Structures of high-density and low-density water. *Phys. Rev. Lett.* **84**, 2881 (2000).

Data availability

The data supporting the findings of this study are available within the paper, and all input files that are necessary to reproduce the reported results are included in Supplementary Information. All data generated for the study are available upon request from the corresponding author, and the MLP for hydrogen constructed here are available at https://github.com/BingqingCheng/MLP-highP-H.

Acknowledgements We are thankful to G. Ackland, H. Geng. and R. Redmer, who shared their AIMD trajectories for us to benchmark the MLP. We thank S. Sorella for providing the VMC training dataset. We acknowledge D. Frenkel, B. Monserrat, M. Casula, A. M. Saitta, R. Helled, G. Carleo and S. Sorella for discussions. B.C. acknowledges funding from the Swiss National Science Foundation (project P2ELP2-184408), resources provided by the Cambridge Tier-2 system funded by EPSRC Tier-2 capital grant EP/P020259/1 and by CSCS under project ID s957. G.M. acknowledges financial support from the Swiss National Science Foundation through grant number 200021-179312. C.J.P. is supported by the Royal Society through a Royal Society Wolfson Research Merit award and the EPSRC through grant EP/P022596/1. M.C. acknowledges funding from the Swiss National Science Foundation (project 200021-182057).

Author contributions B.C., G.M. and M.C. conceptualized the research; B.C., C.J.P. and M.C. performed the research and analysed the data; B.C., G.M., C.J.P. and M.C. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2677-y.
Correspondence and requests for materials should be addressed to B.C.
Peer review information *Nature* thanks Graeme Ackland and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Rare-earth–platinum alloy nanoparticles in mesoporous zeolite for catalysis

https://doi.org/10.1038/s41586-020-2671-4

Received: 1 October 2019

Accepted: 22 June 2020

Published online: 9 September 2020

Check for updates

Ryong Ryoo[1,2]✉, Jaeheon Kim[1], Changbum Jo[1,3], Seung Won Han[1,2], Jeong-Chul Kim[1], Hongjun Park[1,2], Jongho Han[1,2], Hye Sun Shin[1,2,4] & Jae Won Shin[1]

Platinum is a much used catalyst that, in petrochemical processes, is often alloyed with other metals to improve catalytic activity, selectivity and longevity[1–5]. Such catalysts are usually prepared in the form of metallic nanoparticles supported on porous solids, and their production involves reducing metal precursor compounds under a H_2 flow at high temperatures[6]. The method works well when using easily reducible late transition metals, but Pt alloy formation with rare-earth elements through the H_2 reduction route is almost impossible owing to the low chemical potential of rare-earth element oxides[6]. Here we use as support a mesoporous zeolite that has pore walls with surface framework defects (called 'silanol nests') and show that the zeolite enables alloy formation between Pt and rare-earth elements. We find that the silanol nests enable the rare-earth elements to exist as single atomic species with a substantially higher chemical potential compared with that of the bulk oxide, making it possible for them to diffuse onto Pt. High-resolution transmission electron microscopy and hydrogen chemisorption measurements indicate that the resultant bimetallic nanoparticles supported on the mesoporous zeolite are intermetallic compounds, which we find to be stable, highly active and selective catalysts for the propane dehydrogenation reaction. When used with late transition metals, the same preparation strategy produces Pt alloy catalysts that incorporate an unusually large amount of the second metal and, in the case of the PtCo alloy, show high catalytic activity and selectivity in the preferential oxidation of carbon monoxide in H_2.

Propane dehydrogenation (PDH) is one of the most important petrochemical processes, used for a large portion of the worldwide propylene production[3]. This process is currently attracting interest in the area of heterogeneous catalysis owing to a sudden increase in propane supply from the shale gas revolution. The industrial PDH process has been using PtSn bimetallic catalysts supported on porous alumina for nearly three decades, since its discovery in the early 1990s[3,7]. The Pt metal alone exhibits high initial catalytic activity, but rapidly deactivates as a result of coke deposition on the Pt surface. The coke deposition also causes an undesirable loss of the catalytic selectivity to propylene. To alleviate these problems, Sn is introduced to form an alloy with Pt nanoparticles and thereby dilute the Pt surface with inactive Sn atoms and break up coke-generating Pt ensembles. But coke deposition still occurs, and currently used PtSn/alumina catalysts still require frequent and cumbersome regeneration steps to restore their catalytic activity.

Our search for a highly active and more durable PDH catalyst has focused on using a siliceous MFI zeolite with a hierarchical micro-/mesoporous structure as a replacement for the alumina support. This zeolite, synthesized using a multi-ammonium surfactant as a structure-directing agent, has attracted attention as an advanced catalyst support[8–13] owing to its unique structure, which comprises ultrathin zeolite frameworks and three-dimensionally interconnected mesopores that boost the catalytic performance of supported metals by providing both facile reactant and product diffusion and high metal nanoparticle dispersion within the mesopores[12,13]. We incorporated La and Y along with Pt into this mesoporous zeolite to improve the Pt catalyst dispersion, as these rare-earth element (REE) oxides are known for their strong metal–support interactions with Pt that stabilize the latter in the form of small nanoparticles[14,15]. Whereas we expected that addition of REEs might improve the initial PDH activity of the Pt/zeolite catalyst owing to the strong metal–support interaction effect, to our surprise, the catalytic lifetime was increased more than tenfold with the addition of La and Y. Investigation by atomic-resolution electron microscopy showed that some La and Y existed in an alloyed form with Pt nanoparticles, supported on the mesoporous zeolite. Systematic exploration allowed us to increase the catalytic lifetime 700-fold by ensuring that supported Pt nanoparticles are all present as intermetallic compounds with La.

Figure 1 shows the structure of the PtY intermetallic compound nanoparticles formed on the mesoporous zeolite. Representative low-magnification high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) images (Fig. 1a, b) show metal nanoparticles about 3 nm in diameter uniformly distributed over the zeolite support, with energy-dispersive X-ray spectroscopy (EDS)

[1]Center for Nanomaterials and Chemical Reactions, Institute for Basic Science (IBS), Daejeon, Republic of Korea. [2]Department of Chemistry, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. [3]Present address: Department of Chemistry and Chemical Engineering, Inha University, Incheon, Republic of Korea. [4]Present address: Particulate Matter Research Center, Research Institute of Industrial Science and Technology (RIST), Gwangyang-si, Republic of Korea. ✉e-mail: rryoo@kaist.ac.kr

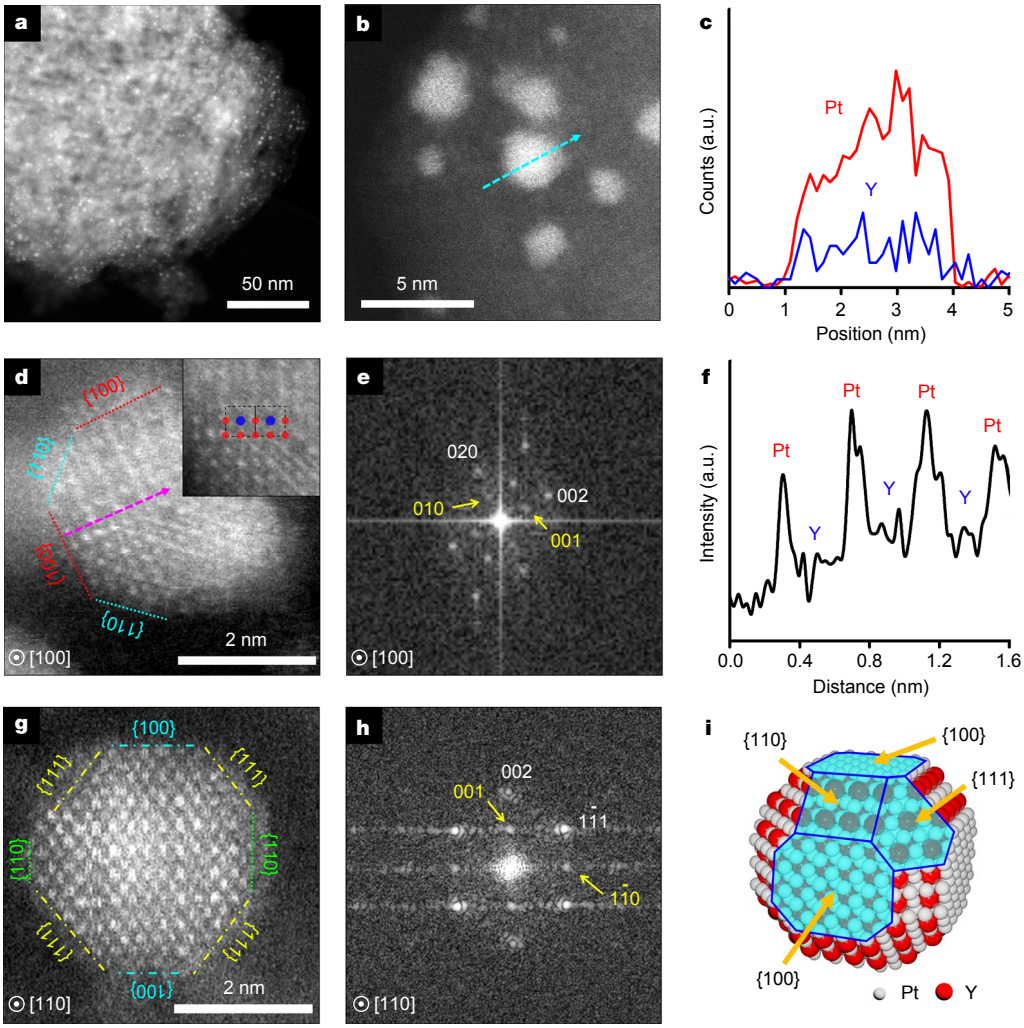


Fig. 1 | Pt_3Y nanoparticles with an $L1_2$ superlattice structure supported on mesoporous MFI zeolite. a, **b**, Low-magnification HAADF-STEM images, showing uniformly sized metal nanoparticles dispersed on the mesoporous zeolite. **c**, EDS line scan profile taken along the blue arrow in **b**, which indicates the presence of both Pt and Y in the same particle. The signals were collected from the Pt M edge and the Y K edge. **d**, **g**, AR-HAADF-STEM images taken along the [100] and [110] zone axes of the metal nanoparticles, showing the Pt_3Y ordered alloy structure with an $L1_2$ superlattice. The inset in **d** is a cropped image showing that each Y column (blue) is surrounded by eight Pt columns (red). **e**, **h**, FFT images obtained from the HAADF-STEM images of **d** and **g**, respectively. $L1_2$ superlattice reflections from the intermetallic compound structure are indicated by yellow arrows. **f**, Intensity profile taken along the ⟨100⟩ direction indicated by a purple arrow in **d**. The intensity profile in **f** shows that Pt and Y atomic columns are alternating. **i**, Proposed atomic structure of the Pt_3Y nanoparticles. Grey and red spheres represent Pt and Y atoms, respectively.

measurements indicating that Pt and Y coexist in the same particle (Fig. 1c). To determine the detailed atomic structure of the bimetallic nanoparticles, atomic-resolution (AR) HAADF-STEM images and the corresponding fast Fourier transform (FFT) images were taken along different zone axes, [100] and [110]. The AR-HAADF-STEM image taken along [100] displays a regular array of bright and dark regions, which correspond to two different types of atomic column aligned perpendicularly to the image (Fig. 1d). The image contrast can be confirmed by the line scan profile shown in Fig. 1f. Considering the Z-contrast of the HAADF detector, the bright parts can be regarded as columns of heavy Pt atoms, whereas the dark parts are assigned to columns of relatively lighter Y atoms. As shown in the inset of Fig. 1d, when projected along [100], every column of Y atoms is neighboured by eight columns of Pt. This ordered atomic arrangement is identical to the $L1_2$ superlattice with an atomic composition of Pt_3Y found in bulk metal alloys, except for the slight difference in the atomic distance due to nanocrystallinity. The Pt_3Y $L1_2$ structure can be described as the replacement of corner Pt atoms in the unit cell of face-centred cubic Pt metal with bulkier Y atoms. The inter-planar distances and angles shown in the AR-HAADF-STEM image are consistent with the Pt_3Y $L1_2$ structure. The (010) and (001) reflection spots marked in yellow in the FFT image (Fig. 1e) are characteristic of the $L1_2$ superlattice. The $L1_2$ arrangement of Pt and Y atoms can also be confirmed when viewed along the [110] zone axis (Fig. 1g, h). On the basis of the STEM investigation, a structural model of the Pt_3Y nanoparticle is proposed in Fig. 1i. The same $L1_2$ superlattice structure was also observed in the case of PtLa nanoparticles supported on mesoporous zeolite. A detailed structural analysis of the Pt_3La nanoparticle is provided in Extended Data Fig. 1.

The Pt–REE incorporation in the mesoporous zeolite was conducted by co-impregnation of aqueous solutions of $Pt(NH_3)_4(NO_3)_2$ and either $La(NO_3)_3{\cdot}6H_2O$ or $Y(NO_3)_3{\cdot}6H_2O$, followed by heating under an O_2 flow at 350 °C and subsequently under H_2 at 700 °C (see Methods). However, the formation of Pt_3La and Pt_3Y intermetallic nanoparticles that exhibited long catalytic lifetime critically depended on the zeolite synthesis procedures, even when the obtained zeolites were all siliceous and had similar mesoporosity. When the zeolite was prepared following a typical

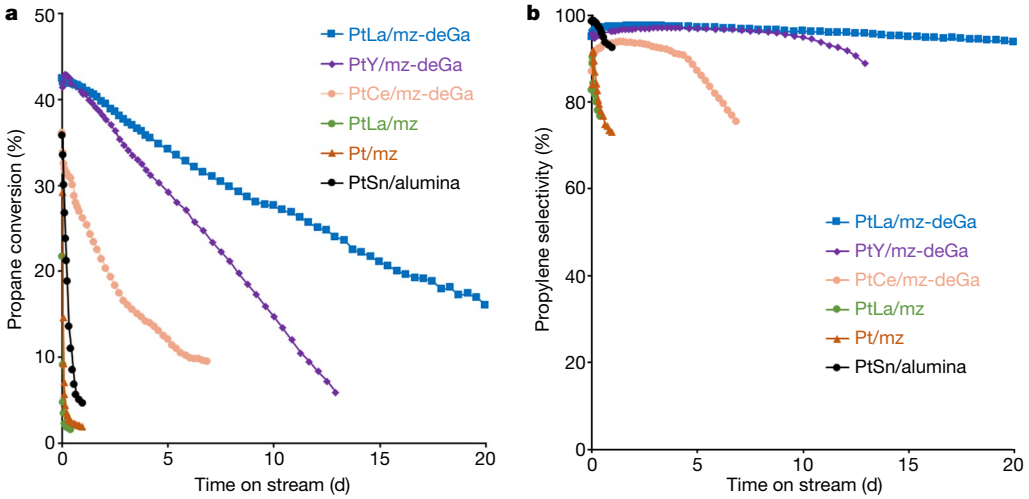


Fig. 2 | Catalytic performance of Pt–REE intermetallic nanoparticles supported on the mesoporous zeolite in propane dehydrogenation. **a**, Propane conversion as a function of time on stream. **b**, Propylene selectivity as a function of time on stream. The bimetallic Pt–REE alloy catalysts are composed of 1 wt% Pt and 1 wt% REE. For comparison, PtSn/alumina and Pt/mz catalysts were also tested. The PDH reaction conditions were as follows: 50 mg catalyst, weight hourly space velocity of 11 h^{-1} with a pure propane gas flow, and temperature of 580 °C.

synthesis procedure using sodium silicate, the impregnated REEs in the zeolite existed as oxide nanoparticles, without forming intermetallic alloy nanoparticles with Pt. The Pt nanoparticles in separation from the REEs exhibited an extremely short PDH catalytic lifetime, identical to that obtained without REE incorporation. Alternatively, when zeolites synthesized under a sodium-free condition were used, there was a small but distinct increase in the catalytic lifetime. Our curiosity about this phenomenon led us to carry out an AR-HAADF-STEM investigation of the Pt–REE/zeolite, in which we observed a tiny portion of supported nanoparticles existing in alloy forms. From this unexpected observation, we speculated that the Pt–REE alloy formation could occur by incorporating the REEs in the form of single atomic species with high chemical potential. Such single atomic REEs could be converted to metallic REEs with H_2 more easily than bulk REE oxides. We further speculated that the formation of atomically dispersed REEs could be assisted by framework defect sites in the zeolite, where multiple silanol groups are adjacently positioned to form a cluster of silanol groups (silanol nest). In fact, silanol nests are often detected by Fourier transform infrared (FT-IR) spectroscopy in mesoporous zeolite synthesized under sodium-free conditions. We believe that the silanol nests could stabilize the single atomic REE species by forming coordination bonds[16]. In that case, it would be possible to control the formation of Pt–REE alloy nanoparticles by the concentration of silanol nests in the zeolite.

With these assumptions, we intentionally generated silanol nests by synthesizing a mesoporous gallosilicate zeolite and subsequently removing the framework Ga atoms using nitric acid (see Methods and Extended Data Figs. 2, 3). Each of the degallated sites corresponded to a generated silanol nest. The increase in the silanol nest concentration in the zeolite was supported by FT–IR measurements (Extended Data Fig. 4). In accordance with our hypothesis, the REEs incorporated in this zeolite were indeed atomically dispersed, as shown in a video taken by AR-HAADF-STEM (see Supplementary Video 1). In this video, the single atomic species of La exhibited rapid and random translational motions on the zeolite surface, which we interpret as a single atomic REE species hopping from one silanol nest to another. We believe that the atomistic diffusion enables the formation of Pt–REE intermetallic nanoparticles when the zeolite is heated under H_2 at 700 °C after impregnation of the metal precursors. In fact, the Pt–REE/zeolite sample shown in Fig. 1 was prepared using a degallated zeolite containing a large number of silanol nest sites. The Pt–REE/zeolite showed formation of Pt_3La and Pt_3Y nanoparticles throughout the zeolite support. The Pt_3La and Pt_3Y nanoparticles generated in this manner exhibited a distinctive H_2 sorption behaviour (see Extended Data Table 1). The alloy nanoparticles chemisorbed 0.5 H atoms per total Pt in the sample in a fully reversible manner, given that all hydrogen completely desorbed upon evacuation for 1 h at room temperature. Monometallic Pt nanoparticles, by contrast, chemisorbed 1.0 H atoms per total Pt, with 64% of this amount strongly and irreversibly bonded. This indicates that the electronic state of the intermetallic compound nanoparticles was markedly different from that of monometallic Pt, as further confirmed by X-ray absorption near-edge structure (XANES) analysis (see Extended Data Fig. 5), which indicated a noticeable shift in edge energy and in the white-line region upon the formation of Pt–REE intermetallic compounds.

The formation of the Pt_3La and Pt_3Y nanoparticles on the degallated zeolite brought a striking enhancement in all aspects of the PDH catalytic performance, including activity, selectivity and durability, as shown in Fig. 2. The catalytic reaction was performed at 580 °C using a pure propane flow at a large space velocity (see Methods), and these harsh operating conditions caused even the conventional PtSn/alumina catalyst to experience severe coke formation and catalyst deactivation within 1 d. In sharp contrast, the Pt_3La intermetallic nanoparticles supported on the degallated zeolite (denoted as PtLa/mz-deGa, where mz stands for mesoporous zeolite) showed a high initial propane conversion of 40% (close to the equilibrium conversion) and underwent extremely slow deactivation, retaining 8% conversion even after 30 days of reaction. In a control using a catalyst with the same amounts of Pt and La supported on the mesoporous zeolite synthesized from sodium silicate (denoted as PtLa/mz), propane conversion was 22% initially and dropped very rapidly to below 5% within 1.5 h. Even though both catalysts used zeolites with the same MFI framework structure and the same mesoporosity as the support (Extended Data Fig. 6), their catalytic lifetimes differed by three orders of magnitude. The rapid deactivation behaviour of the PtLa/mz catalyst was similar to that of the monometallic Pt catalyst supported on mesoporous zeolite (Pt/mz), consistent with the hydrogen chemisorption property of the PtLa/mz sample, indicating that the supported Pt existed entirely as monometallic nanoparticles and that no alloy with La was present (Extended Data Table 1). These comparisons suggest that the differences in catalyst performance between PtLa/mz-deGa and PtLa/mz are largely due to formation of the intermetallic alloy. Nevertheless, it is worth noting that the PtLa/mz-deGa catalyst contained a small amount of Ga (Si/Ga = 598) even after the degallation, and that this may

participate in the catalytic function[17,18]. The effect of the remaining Ga was examined from the differences in PDH performance of mz-deGa, La/mz-deGa and Pt/mz-deGa, as shown in Extended Data Fig. 7. A simple comparison of the catalytic result suggested that the Ga effect was not as dramatic as the alloying effect, but there are still possibilities of Ga acting somehow as a promotor to enhance the catalytic function and/or the alloy formation in PtLa/mz-deGa. Moreover, in a similar manner to PtLa/mz-deGa, PtY/mz-deGa also exhibited high initial propane conversion and propylene selectivity, and slow catalyst deactivation. This appears to be attributable to the difference in atomic size and electron negativity. Because La has a larger atomic size and lower electron negativity than Y, the geometric and electronic properties of Pt were modified to a greater extent. As this result shows, zeolites can be an effective support for the development of PDH catalysts. In recent years, efficient conversion of propane to propylene was reported with PtZn and PtSn catalysts supported on zeolites[19–21]. For industrial application of the zeolite-supported catalysts, it would be important to establish an efficient regeneration method, such as the oxychlorination process employed for PtSn/alumina. However, the regenerability of these catalysts has not yet been fully demonstrated.

The high PDH catalytic performance of PtLa/mz-deGa and PtY/mz-deGa can be ascribed to the critical role of surface silanol nests that enabled atomistic alloying of La and Y into Pt nanoparticles. This approach can be extended to produce other metal alloys composed of Pt-group metals and REEs, as suggested in the case of the PtCe intermetallic alloy nanoparticle catalyst shown in Fig. 2 and Extended Data Fig. 8. The silanol nests in the mesoporous zeolite can effectively activate the hard-to-reduce REEs (that is, REEs with low reduction potential) to overcome the energy hurdle by putting them into single atomic species, which facilitates the formation of alloy nanoparticles. For the PtLa and PtY bimetallic catalysts, the reduction potential of the REEs is so low that the role of silanol nests is decisive for the formation of the metal alloys. In many other transition metal elements with higher reduction potential (for example, Co, Fe and Zn), alloy formation is known to occur spontaneously when the incorporated metal precursors are heated under H_2. But even in this case, the atomistic diffusivity of the activated species via silanol nests is expected to facilitate access of the transition metal elements to nearby Pt nanoparticles, on which these species are reduced to the metallic state by reacting with chemisorbed hydrogen. This process can promote metal incorporation into the Pt nanoparticle. In fact, the promoting effect has been confirmed in the case of PtCo bimetallic alloy catalysts for the preferential CO oxidation (PROX) reaction. The PtCo catalysts were prepared on two zeolite supports, mz and mz-deGa, and on an alumina support. All the resultant catalyst samples showed PtCo nanoparticles with randomly distributed Pt and Co atoms, but the Co/Pt atomic ratio was considerably different, increasing in the order of alumina ≪ mz < mz-deGa. The difference in the alloy composition resulted in a remarkable change in the catalytic performance for PROX in a H_2-rich stream containing 1.4% CO (catalytic efficiency: PtCo/alumina ≪ PtCo/mz < PtCo/mz-deGa; see Extended Data Fig. 9). On the basis of these results, we believe that the use of the present mesoporous zeolite as a support would allow the discovery of not only REE-based alloy catalysts, but also other transition metal-based alloy catalysts with various compositions and structures, opening up new opportunities for catalytic applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2671-4.

1. Yu, W., Porosoff, M. D. & Chen, J. G. Review of Pt-based bimetallic catalysts: from model surfaces to supported catalysts. *Chem. Rev.* **112**, 5780–5817 (2012).
2. Bariås, O. A., Holmen, A. & Blekkan, E. A. Propane dehydrogenation over supported Pt and Pt-Sn catalysts: II. Catalyst preparation, characterization, and activity measurements. *J. Catal.* **158**, 1–12 (1996).
3. Sattler, J. J. H. B., Ruiz-Martinez, J., Santillan-Jimenez, E. & Weckhuysen, B. M. Catalytic dehydrogenation of light alkanes on metals and metal oxides. *Chem. Rev.* **114**, 10613–10653 (2014).
4. Burch, R. & Garla, L. C. Platinum-tin reforming catalysts: activity and selectivity in hydrocarbon reactions. *J. Catal.* **71**, 360–372 (1981).
5. Liu, K., Wang, A. & Zhang, T. Recent advances in preferential oxidation of CO reaction over platinum group metal catalysts. *ACS Catal.* **2**, 1165–1178 (2012).
6. Furukawa, S. & Komatsu, T. Intermetallic compounds: promising inorganic materials for well-structured and electronically modified reaction environments for efficient catalysis. *ACS Catal.* **7**, 735–765 (2017).
7. Buyanov, R. A. & Pakhomov, N. A. Catalysts and processes for paraffin and olefin dehydrogenation. *Kinet. Catal.* **42**, 64–75 (2001).
8. Choi, M., Na, K., Kim, J. & Ryoo, R. Stable single-unit-cell nanosheets of zeolite MFI as active and long-lived catalysts. *Nature* **461**, 246–249 (2009).
9. Na, K. et al. Directing zeolite structures into hierarchically nanoporous architectures. *Science* **333**, 328–332 (2011).
10. Jo, C., Cho, K., Kim, J. & Ryoo, R. MFI zeolite nanosponges possessing uniform mesopores generated by bulk crystal seeding in the hierarchical surfactant-directed synthesis. *Chem. Commun.* **50**, 4175–4177 (2014).
11. Lee, S., Jo, C. & Ryoo, R. Tomographic imaging of pore networks and connectivity of surfactant-directed mesoporous zeolites. *J. Mater. Chem. A* **5**, 11086–11093 (2017).
12. Kim, J., Han, S., Kim, J.-C. & Ryoo, R. Supporting nickel to replace platinum on zeolite nanosponges for catalytic hydroisomerization of *n*-dodecane. *ACS Catal.* **8**, 10545–10554 (2018).
13. Kim, J.-C. et al. Mesoporous MFI zeolite nanosponge supporting cobalt nanoparticles as a Fischer–Tropsch catalyst with high yield of branched hydrocarbons in the gasoline range. *ACS Catal.* **4**, 3919–3927 (2014).
14. Meriaudeau, P., Dutel, J. F., Dufax, M. & Naccache, C. Further investigation on metal-support interaction: TiO_2, CeO_2, SiO_2 supported platinum catalysts. *Stud. Surf. Sci. Catal.* **11**, 95–104 (1982).
15. Fleisch, T. H., Hicks, R. F. & Bell, A. T. An XPS study of metal-support interactions on $PdSiO_2$ and $PdLa_2O_3$. *J. Catal.* **87**, 398–413 (1984).
16. Deng, C. et al. The effect of positioning cations on acidity and stability of the framework structure of Y zeolite. *Sci. Rep.* **6**, 23382 (2016).
17. Sattler, J. J. H. B. et al. Platinum-promoted Ga/Al_2O_3 as highly active, selective, and stable catalyst for the dehydrogenation of propane. *Angew. Chem. Int. Ed.* **53**, 9251–9256 (2014).
18. Choi, S. W. et al. Propane dehydrogenation catalyzed by gallosilicate MFI zeolites with perturbed acidity. *J. Catal.* **345**, 113–123 (2017).
19. Sun, Q. et al. Subnanometer bimetallic platinum–zinc clusters in zeolites for propane dehydrogenation. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.202003349 (2020).
20. Xu, Z., Yue, Y., Bao, X., Xie, Z. & Zhu, H. Propane dehydrogenation over Pt clusters localized at the Sn single-site in zeolite framework. *ACS Catal.* **10**, 818–828 (2020).
21. Liu, L. et al. Regioselective generation and reactivity control of subnanometric platinum clusters in zeolites for high-temperature catalysis. *Nat. Mater.* **18**, 866–873 (2019).

Methods

Preparation of zeolites

The mz-deGa zeolite was prepared by the post-synthetic degallation of mesoporous MFI gallosilicate zeolite, which was prepared using $[C_{18}H_{37}–N^+(Me)_2–C_6H_{12}–N^+(Me)_2–C_4H_9]Br^-_2$ (denoted as '$C_{18\text{-}6\text{-}4}$') as a micro-/mesoporous structure-directing agent. For the synthesis of mesoporous MFI gallosilicate, sodium silicate (29 wt% SiO_2, Si/Na = 1.75; Shinheung Silicate) and gallium nitrate were used as silica and gallium oxide sources, respectively. In the typical synthesis of mesoporous MFI gallosilicate, 20 g of sodium silicate solution was diluted with 24 g of distilled water. The diluted sodium silicate solution was mixed with a clear solution consisting of 36 g of water and 4.7 g of $C_{18\text{-}6\text{-}4}$. After ageing under magnetic stirring for 30 min, the clear solution containing 0.52 g of gallium nitrate and 15.4 g of water was poured at once into the solution containing sodium silicate and $C_{18\text{-}6\text{-}4}$. The resultant gel was aged at 60 °C for 6 h under magnetic stirring. After cooling to the ambient temperature, 18.7 g of 0.86 M H_2SO_4 was added dropwise with vigorous stirring to adjust the pH of the zeolite synthesis gel. Again, the resultant gel was aged under magnetic stirring at 60 °C for 12 h. The final molar gel composition was SiO_2:Ga_2O_3:$C_{18\text{-}6\text{-}4}$:Na_2O:H_2SO_4:H_2O = 100:1:7.5:30:16:6,000. The resultant gel was transferred into a Teflon-lined stainless-steel autoclave and then heated by tumbling at 150 °C for 3 d in a convection oven. The product was collected as a white powder by filtration, washing with distilled water and drying at 100 °C. Finally, mesoporous MFI gallosilicate was obtained after the calcination at 580 °C to remove the structure-directing agent.

The framework Ga species in the mesoporous MFI gallosilicate was removed by HNO_3 treatments. In a typical Ga-removal procedure, 1 g of mesoporous MFI gallosilicate was poured into 100 ml of the 13 M HNO_3. The resultant mixture was heated at 100 °C for 12 h under magnetic stirring. After filtration, the resultant sample was thoroughly washed with distilled water until the pH of the filtrate water reached 7. The HNO_3 treatment and H_2O washing were repeated two more times, and then the mz-deGa was collected after drying at 100 °C.

The mz zeolite was synthesized by a procedure similar to that described for the mesoporous MFI gallosilicate. In a typical synthesis of mz, 20 g of sodium silicate (29 wt% SiO_2, Si/Na = 1.75; Shinheung Silicate) was diluted with 24 g of distilled water. To this silicate solution, a clear aqueous solution containing 4.7 g of $C_{18\text{-}6\text{-}4}$ and 67 g of distilled water was poured at once and strongly shaken manually for 10 min. The resultant gel was aged at 60 °C for 6 h under magnetic stirring. The final molar gel composition was SiO_2:$C_{18\text{-}6\text{-}4}$:Na_2O:H_2O = 100:7.5:10:6,000. The final synthesis gel was transferred into a Teflon-lined stainless-steel autoclave and heated by tumbling at 150 °C for 2.5 d. Then, the same procedure was used as for the mesoporous MFI gallosilicate.

Preparation of supported metal catalysts

The Pt and REEs were supported on the mz and mz-deGa zeolites by the incipient wetness impregnation technique. $Pt(NH_3)_4(NO_3)_2$, $La(NO_3)_3{\cdot}6H_2O$, $Y(NO_3)_3{\cdot}6H_2O$ and $Ce(NO_3)_3{\cdot}6H_2O$ were purchased from Sigma-Aldrich and used as received as metal precursors. Typically, Pt and REE precursors were dissolved in an appropriate volume of distilled water, and their amounts were determined so as to yield metal contents of 1 wt% Pt and 1 wt% REE in the final supported catalyst. The Pt–REE-impregnated zeolites were dried at 60 °C overnight. The dried samples were treated under an O_2 flow at 350 °C for 2 h (ramping rate, 0.8 °C min^{-1}; flow rate, 500 $cm^3\,min^{-1}\,g_{cat}{}^{-1}$, where g_{cat} denotes grams of catalyst). Subsequently, the resultant samples were treated under a H_2 flow at 700 °C for 2 h, except for the case of PtCe/mz-deGa, which was treated at 580 °C (ramping rate, 0.3 °C min^{-1}, H_2 flow rate, 300 $cm^3\,min^{-1}\,g_{cat}{}^{-1}$). These zeolite-supported Pt–REE samples were used as catalysts for the PDH reaction tests and for various characterizations, except for the AR-HAADF-STEM investigation.

For the AR-HAADF-STEM investigation, the Pt–REE/zeolite samples were prepared following the aforementioned co-impregnation procedure, except that the Pt and REE loading amounts were changed to 4 wt% and 4 wt%, respectively, in order to increase the average size of supported nanoparticles to obtain clearer images. In addition, Pt/mz and PtSn/alumina were prepared as described in the legend of Extended Data Fig. 10 and tested as catalysts for the PDH reaction. The Pt–Co catalysts were supported onto the mz-deGa, mz and alumina supports following a similar procedure to that for the supported Pt–REE catalysts for the PROX tests. In a typical PtCo loading process, $Pt(NH_3)_4(NO_3)_2$ and $Co(NO_3)_2$ precursors were co-incorporated onto the supports via the incipient wetness impregnation technique, and the amounts of metal precursors were determined to yield 2 wt% and 1 wt% of Pt and Co, respectively, in the final supported PtCo samples. The metal precursor-incorporated samples were heated under an O_2 flow at 350 °C for 2 h (ramping rate of 1.5 °C min^{-1}) and subsequently treated under a H_2 flow at 300 °C for 2 h (ramping rate of 1.3 °C min^{-1}).

Characterizations

The HAADF-STEM images were taken with a Titan cubed G2, a Titan Themis Z and a Titan ETEM G2 instrument at 300 kV acceleration voltage. EDS analysis was carried out with four integrated silicon-drift EDS detectors (Super-X) at a collection solid angle of 0.7 srad. The FT-IR spectra were collected using a JASCO FT-IR 6100 spectrometer under vacuum at room temperature after self-supporting wafers of mesoporous zeolites were degassed at 400 °C. The hydrogen chemisorption was measured using a laboratory-made volumetric apparatus. Prior to the hydrogen chemisorption measurement, supported-metal samples were reduced under a H_2 flow at the appropriate temperature for 2 h. Subsequently, the H_2-treated samples were heated at 400 °C under vacuum for the desorption of hydrogen. After cooling to room temperature, the first hydrogen chemisorption measurement was performed with the degassed sample at room temperature. After degassing the sample again by evacuation at room temperature for 1 h, the second hydrogen chemisorption measurement was carried out. The hydrogen per metal atom in the chemisorption measurement was calculated by extrapolation of the adsorption isotherm to zero pressure. The elemental contents were determined by inductively coupled plasma–atomic emission spectroscopy using an OPTIMA 4300 DV instrument (PerkinElmer). Powder X-ray diffraction (PXRD) patterns were collected by a Rigaku Multiflex diffractometer using a nickel-filtered Cu Kα radiation beam (40 kV, 30 mA). Nitrogen adsorption–desorption isotherms were measured at liquid-nitrogen temperature with a Micromeritics Tristar instrument. The specific surface area was calculated using the Brunauer–Emmett–Teller (BET) equation with the adsorption isotherm data in the relative-pressure range P/P_0 = 0.05–0.2. The pore size distribution was determined using the Barret–Joyner–Halenda (BJH) method from the adsorption branch of the isotherm. The total pore volume was obtained at P/P_0 = 0.95 from the adsorption branch of the isotherm. The XANES spectra were measured at the Pt L_3 edge and the Y K edge in transmission mode at Pohang Accelerator Laboratory (10C-Wide XAFS beamline). For the XANES measurement, each metal precursor-impregnated sample was treated by O_2 at 350 °C and then pressed into a round-shaped pellet (diameter, 10 mm). The O_2-treated pellet was loaded inside a laboratory-made Pyrex glass apparatus and heated under a H_2 flow at the appropriate temperature. After the H_2 treatment, the catalyst pellet was sealed under a H_2 flow inside the Pyrex glass with polyimide windows by melting the glass with a propane torch. The XANES spectra were measured by passing the beam through the glass-sealed catalyst pellets. All the XANES spectra were calibrated using Pt and Y metal foil references, which were purchased from Alfa Aesar. All the analysis of XANES data was performed using ATHENA software[22].

PDH reaction tests

PDH reaction tests were carried out using a fixed-bed reactor made of a quartz tube with an inner diameter of 8 mm. Prior to the catalyst loading, the supported metal catalysts were shaped into particles with diameters of 150–350 μm by pressing the powder form of the catalysts and subsequently sieving with a mesh network. 50 mg of the shaped catalyst was loaded in the quartz tube reactor, and the residual space in the reactor was also filled with quartz sand to minimize the contribution of the thermal reaction. After the loading, the catalysts were activated by in situ reduction under a H_2 flow at the appropriate temperatures for the metal elements (700 °C for PtLa and PtY; 580 °C for Pt and PtCe) for 2 h at a ramping rate of 0.3 °C min^{-1} and a flow rate of 200 $cm^3\ min^{-1}\ g_{cat}^{-1}$. The reduced catalyst bed was then purged with a N_2 gas flow (200 $cm^3\ min^{-1}\ g_{cat}^{-1}$) for 1 h at 580 °C to completely remove the chemisorbed hydrogens. After the N_2 purging, the PDH reaction test was performed with the following reaction conditions: 50 mg catalyst, weight hourly space velocity (WHSV; based on the total weight of catalyst) of 11 h^{-1} with pure propane gas flow and temperature of 580 °C. The products were analysed by an on-line gas chromatography instrument equipped with a flame-ionization detector (FID) and a GS-Gaspro column. The propane conversion and propylene selectivity were calculated on a carbon basis, which were determined by the FID peak areas of the outlet product stream—that is, propane conversion (%) = 100 × (1 − carbons of propane/total carbons); propylene selectivity (%) = 100 × carbons of propylene/(total carbons − carbons of propane). After the PDH test, the quartz sands and catalyst particles loaded in the reactor were collected and investigated by thermogravimetric analysis to determine the amount of coke deposit. The result indicated that the coke formation was negligible, and the product distribution could be entirely determined from the FID peaks.

PROX tests

The PROX tests were performed using a gas-flow fixed-bed reactor made of a quartz tube with 8 mm inner diameter. Prior to the catalyst loading, the PtCo catalysts were shaped in the same manner as in the PDH reaction tests. The shaped catalyst (50 mg) was loaded into the quartz reactor and reduced under a H_2 flow (600 $cm^3\ min^{-1}\ g_{cat}^{-1}$) at 300 °C for 2 h. After cooling the reactor to room temperature under an inert gas flow (600 $cm^3\ min^{-1}\ g_{cat}^{-1}$), the PROX test was performed with the following reaction conditions: 50 mg catalyst; molar gas composition $CO:O_2:H_2:He:N_2$ = 1.4:1.4:56.8:39.0:1.4; GHSV = 36,000 $cm^3\ h^{-1}$; and reaction temperature = room temperature–225 °C. The products were analysed with an on-line gas chromatography instrument equipped with an FID and a thermal conductivity detector using a Porapak Q column and a Molecular Sieve 5A column, respectively.

Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

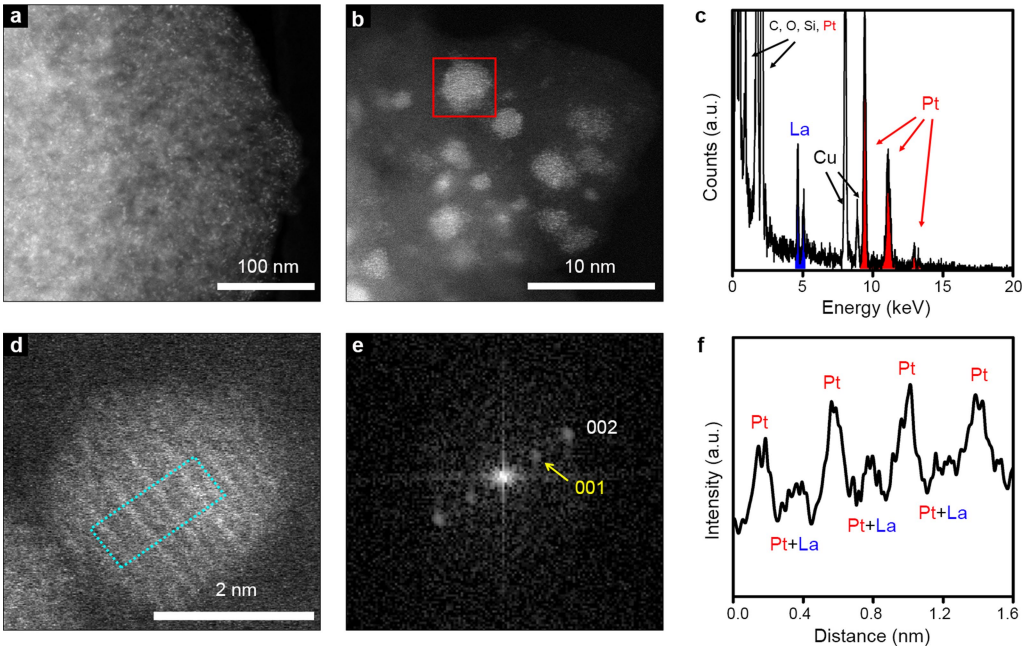
22. Ravel, B. & Newville, M. ATHENA and ARTEMIS: interactive graphical data analysis using IFEFFIT. *Phys. Scr.* **2005**, T115 (2005).

Acknowledgements This work was supported by IBS-R004-D1. We thank M. G. Kim at Pohang Accelerator Laboratory (PAL) for allocating special beam time for XAFS measurements. Electron microscopy experiments were performed with help from H. B. Bae and J. S. Choi at KAIST Analysis Center for Research Advancement (KARA) and C. S. Kim at Korea Institute of Ceramic Engineering and Technology (KICET).

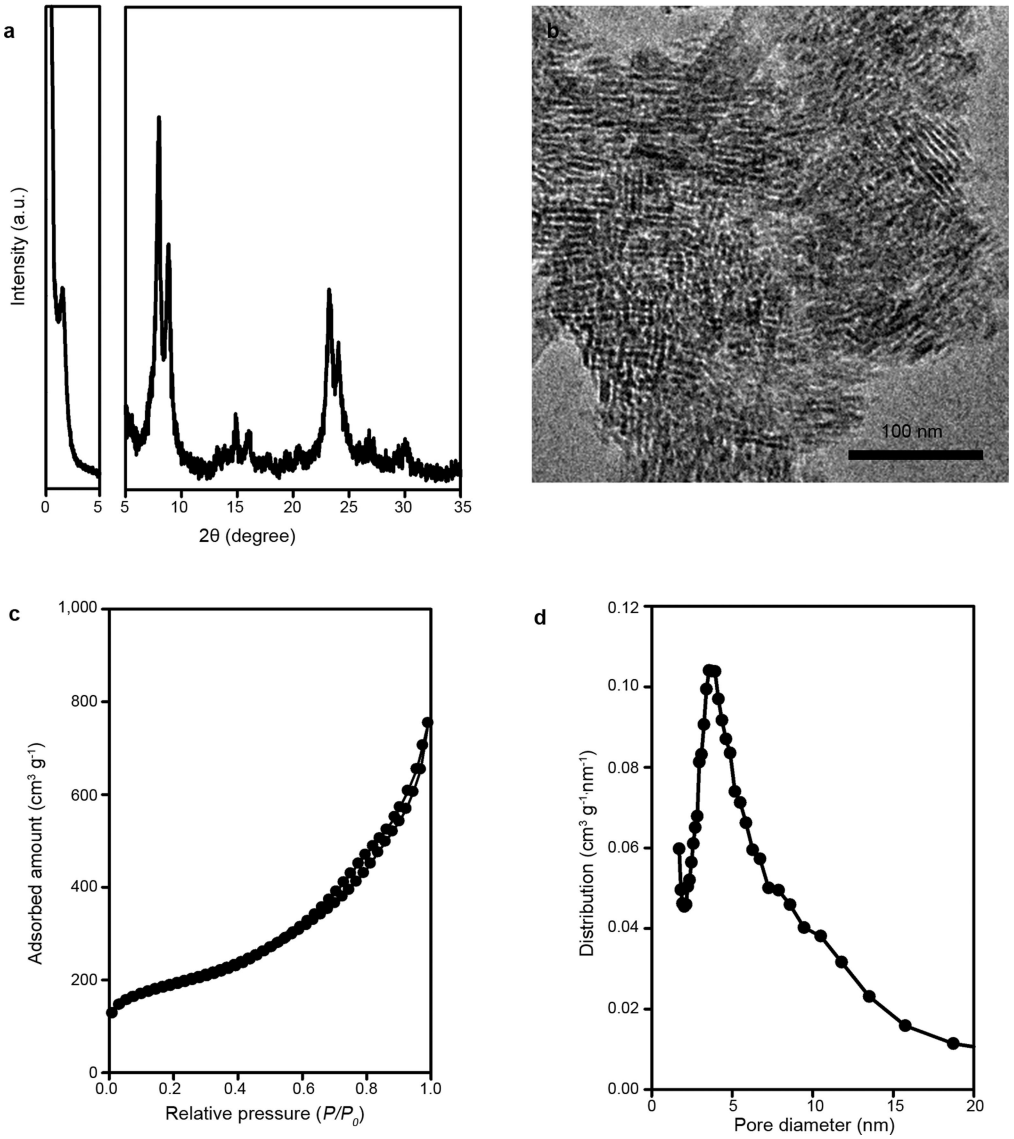
Author contributions R.R. initiated and led the project. R.R., J.K. and S.W.H. wrote the manuscript. J.K., C.J., S.W.H., J.-C.K., H.P., J.H. and H.S.S. carried out materials synthesis, characterization and catalytic measurements. J.W.S. and J.-C.K. performed the TEM investigation.

Competing interests The authors declare no competing interests.

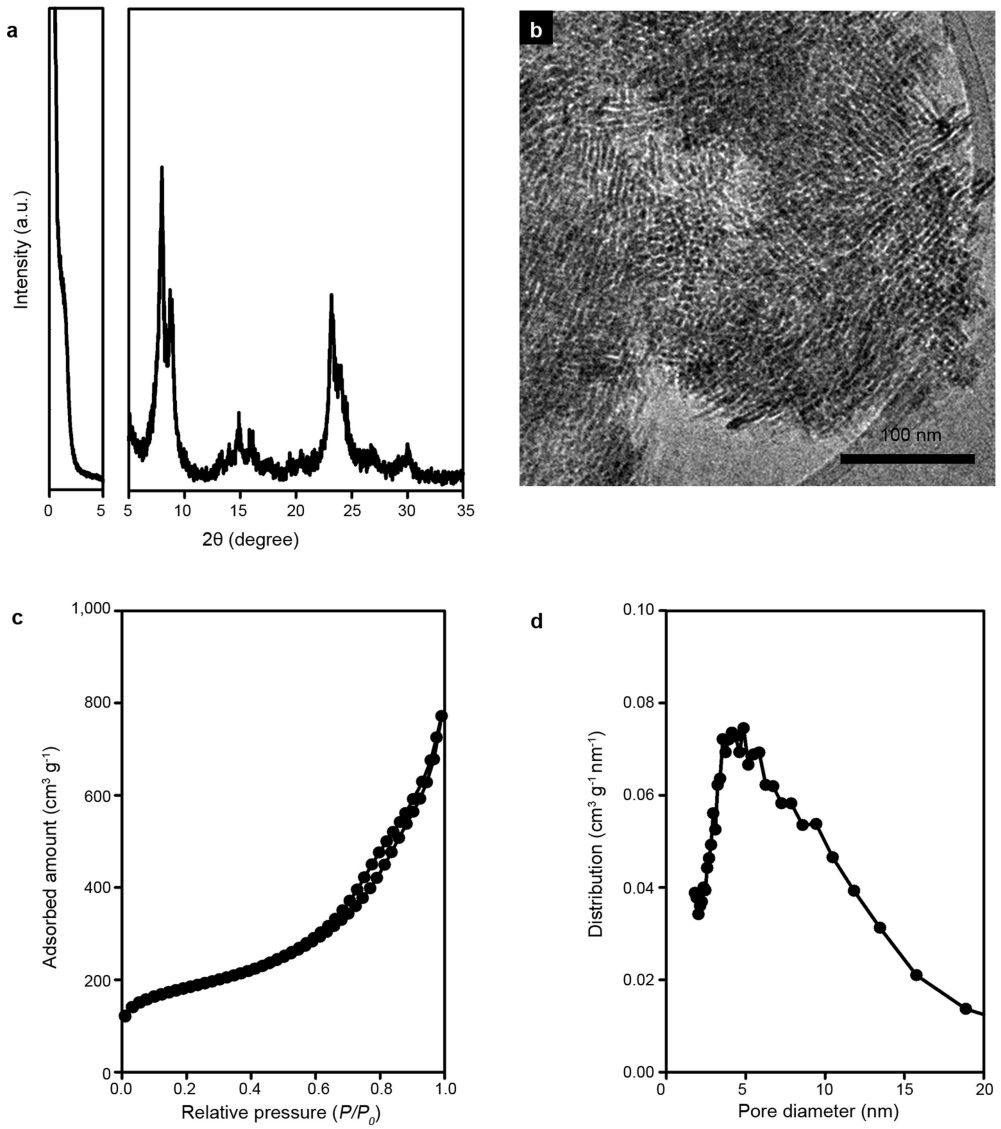
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2671-4.
Correspondence and requests for materials should be addressed to R.R.



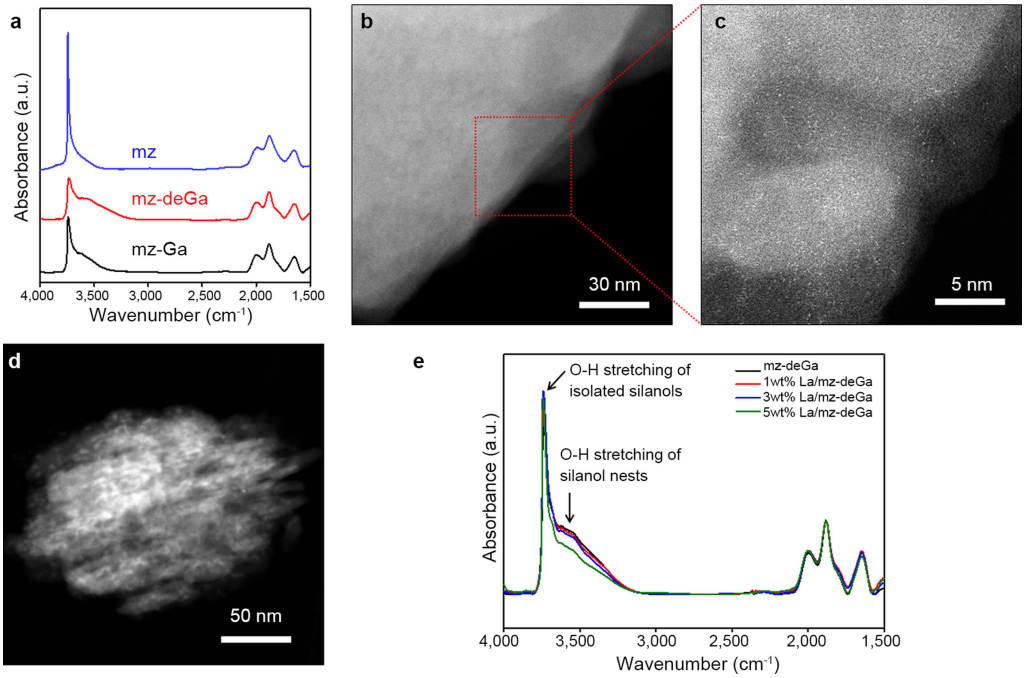
Extended Data Fig. 1 | Pt_3La nanoparticles with an $L1_2$ superlattice structure supported on mz-deGa. a, **b**, Low-magnification HAADF-STEM images showing uniformly sized metal nanoparticles dispersed on the mesoporous zeolites. **c**, EDS spectrum taken from the red box in **b**, indicating the presence of both Pt and La in the same particle. **d**, AR-HAADF-STEM image of a metal nanoparticle, showing the Pt_3La ordered alloy structure with the $L1_2$ superlattice. **e**, FFT image from the HAADF-STEM image of **d**. The $L1_2$ superlattice reflection from the intermetallic compound structure is indicated by a yellow arrow. **f**, Intensity profile taken along the light-blue box in **d**. The intensity profile in **f** shows that Pt and Pt + La atomic columns are alternating.



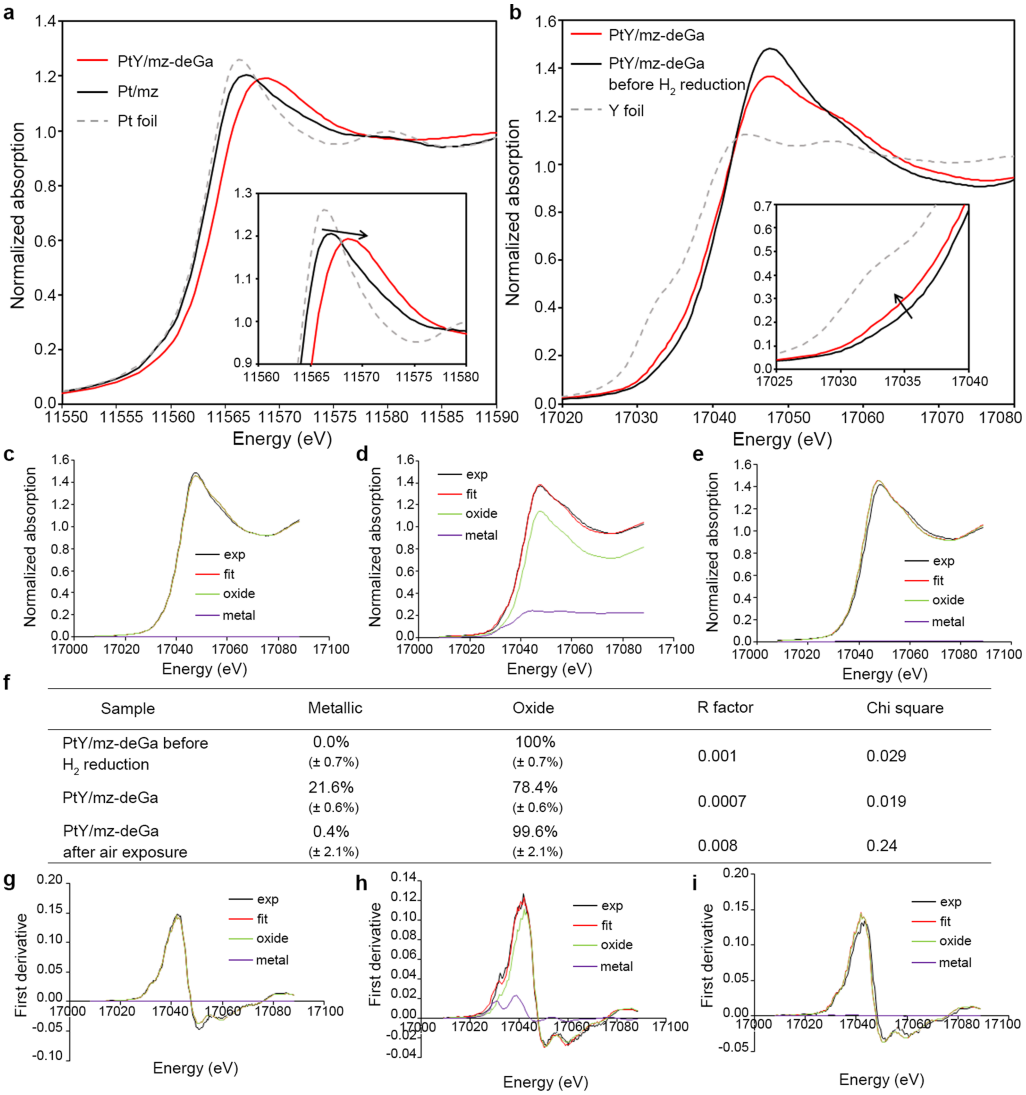
Extended Data Fig. 2 | Structure of mesoporous MFI gallosilicate synthesized using $C_{18\text{-}6\text{-}4}$ as a dual micro-/mesopore structure-directing agent. a, PXRD patterns. **b**, TEM image. **c**, N_2 adsorption–desorption isotherm. **d**, BJH pore size distribution derived from the adsorption branch of the N_2 isotherm. The mesoporous MFI gallosilicate (mz-Ga; parent zeolite of mz-deGa) exhibited highly mesoporous frameworks built with ultrathin zeolitic walls and a uniform mesopore size distribution centred at about 4 nm.



Extended Data Fig. 3 | Structure of mz-deGa obtained by degallation of the mesoporous MFI gallosilicate synthesized using $C_{18\text{-}6\text{-}4}$ as a dual micro-/mesopore structure-directing agent. a, PXRD patterns. **b**, TEM image. **c**, N_2 adsorption–desorption isotherm. **d**, BJH pore size distribution derived from the adsorption branch of the N_2 isotherm. mz-deGa showed similar structural properties to those of the parent mesoporous MFI gallosilicate.



Extended Data Fig. 4 | Atomically dispersed La on mesoporous zeolite with silanol nests. a, FT-IR spectra of mz, mz-deGa and mz-Ga. The mz-Ga sample shows the FT-IR absorbance band corresponding to isolated Si–OH (about 3,750 cm^{-1}) and Ga–OH (3,600 cm^{-1}). The mz-deGa sample displays increased broad FT-IR adsorption at around 3,500 cm^{-1}, which is assigned to the silanol nests. The mz sample shows one sharp FT-IR peak corresponding to the isolated silanols. **b**, HAADF-STEM image of LaO_x-supported mz-deGa showing no noticeable white dots on the grey zeolite matrix, which can be regarded as LaO_x nanoparticles. **c**, AR-HAADF-STEM image of LaO_x/mz-deGa magnified from the STEM image of **b**, revealing the single-atom-dispersed La species as white dots. **d**, HAADF-STEM image of LaO_x-supported mz, showing the nanoparticle-like LaO_x species as white dots. As shown in the HAADF-STEM images and FT-IR spectra, La species can be single-atom-dispersed only if the mesoporous MFI zeolite has sufficient silanol nests. **e**, FT-IR spectra of La-supported mz-deGa samples with various La loadings. The samples were prepared by incipient wetness impregnation of La nitrate and subsequent heat treatment at 350 °C under an O_2 flow. The absorption band corresponding to the silanol nest (~3,500 cm^{-1}; marked with an arrow) was gradually weakened with increasing La loading, indicating that the incorporated La species resulted in the formation of bonds with the silanol nests present in the mz-deGa support.



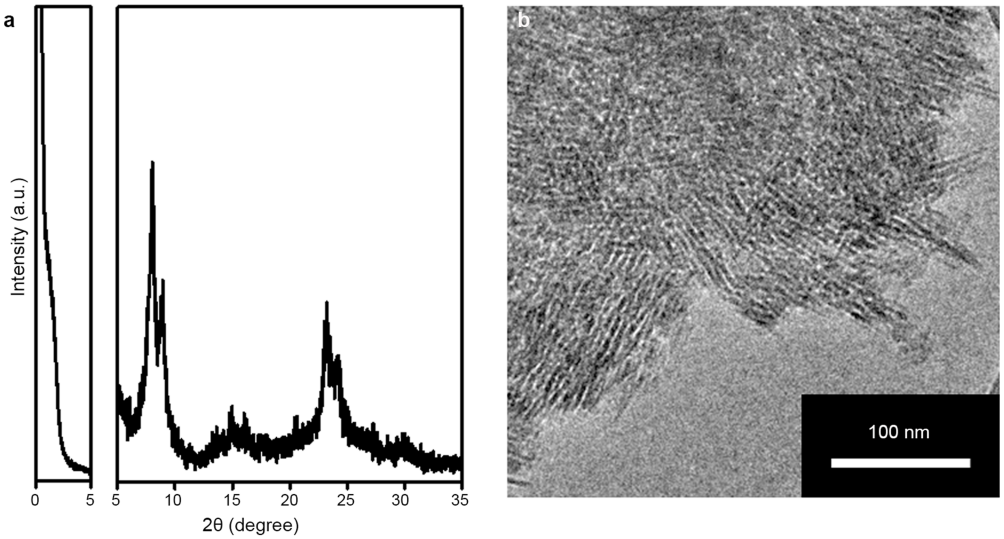
f

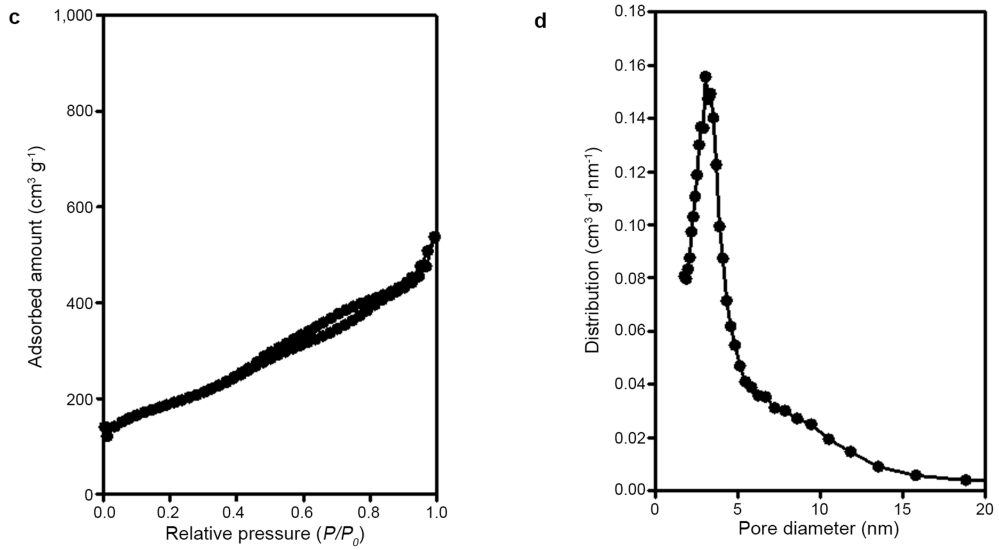
Sample	Metallic	Oxide	R factor	Chi square
PtY/mz-deGa before H_2 reduction	0.0% (± 0.7%)	100% (± 0.7%)	0.001	0.029
PtY/mz-deGa	21.6% (± 0.6%)	78.4% (± 0.6%)	0.0007	0.019
PtY/mz-deGa after air exposure	0.4% (± 2.1%)	99.6% (± 2.1%)	0.008	0.24

j

Sample	Metallic	Oxide	R factor	Chi square
PtY/mz-deGa before H_2 reduction	0.0% (± 0.9%)	100% (± 0.9%)	0.005	0.021
PtY/mz-deGa	24.3% (± 0.8%)	75.7% (± 0.8%)	0.005	0.0015
PtY/mz-deGa after air exposure	1.6% (± 2.1%)	98.4% (± 2.1%)	0.03	0.010

Extended Data Fig. 5 | XANES analysis of PtY/mz-deGa. a, XANES spectra at the Pt L_3 edge. The XANES spectrum of PtY/mz-deGa at the Pt L_3 edge is compared with two reference samples of Pt metal foil and Pt/mz. PtY/mz-deGa showed a notable shift in edge energy and white-line region to higher energy compared to the monometallic Pt reference samples. This indicates that electron donation occurred from less electronegative adjacent Y metals to Pt metals in the bimetallic alloy nanoparticles of PtY/mz-deGa. **b**, XANES spectra at the Y K edge. The PtY/mz-deGa sample is compared with PtY/mz-deGa before H_2 reduction and an Y foil as reference samples. The Y XANES spectrum of PtY/mz-deGa exhibited a noticeable shift in edge energy to lower energy after the reduction. This indicates that a sizable portion of oxidic Y was reduced to metallic Y. **c**–**j**, Linear-fit analysis of the Y K-edge XANES spectra of PtY/mz-deGa and related samples: PtY/mz-deGa before H_2 reduction (that is, sample loaded with metal precursors and calcined with O_2 at 350 °C) at 700 °C (**c**, **g**); PtY/mz-deGa (**d**, **h**); and PtY/mz-deGa after exposure to air (**e**, **i**). In **c**–**e**, the Y K-edge spectra and their fit results are shown. **f** shows the metallic and oxidic Y contents determined by linear fitting of the Y K-edge XANES spectra. In **g**–**i**, the first derivatives of the Y XANES spectra and their fit results are shown. The deconvolution was performed using the Y XANES spectra of two reference samples: (1) Y metal foil and (2) Y_2O_3/mz, which was prepared to have 1 wt% Y by impregnation of yttrium nitrate and subsequent O_2 calcination at 350 °C. **j** shows the metallic and oxidic Y contents, determined by linear fitting of first derivatives of the Y K-edge XANES spectra.



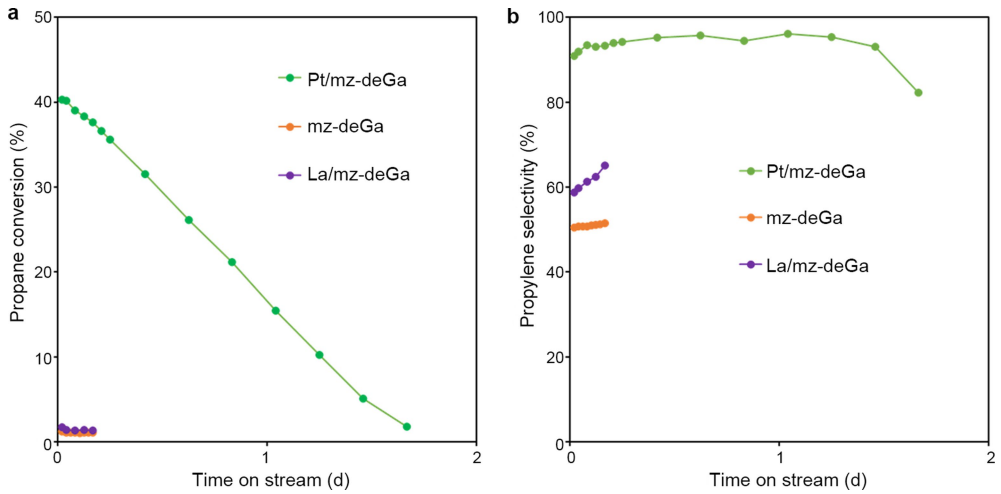


e

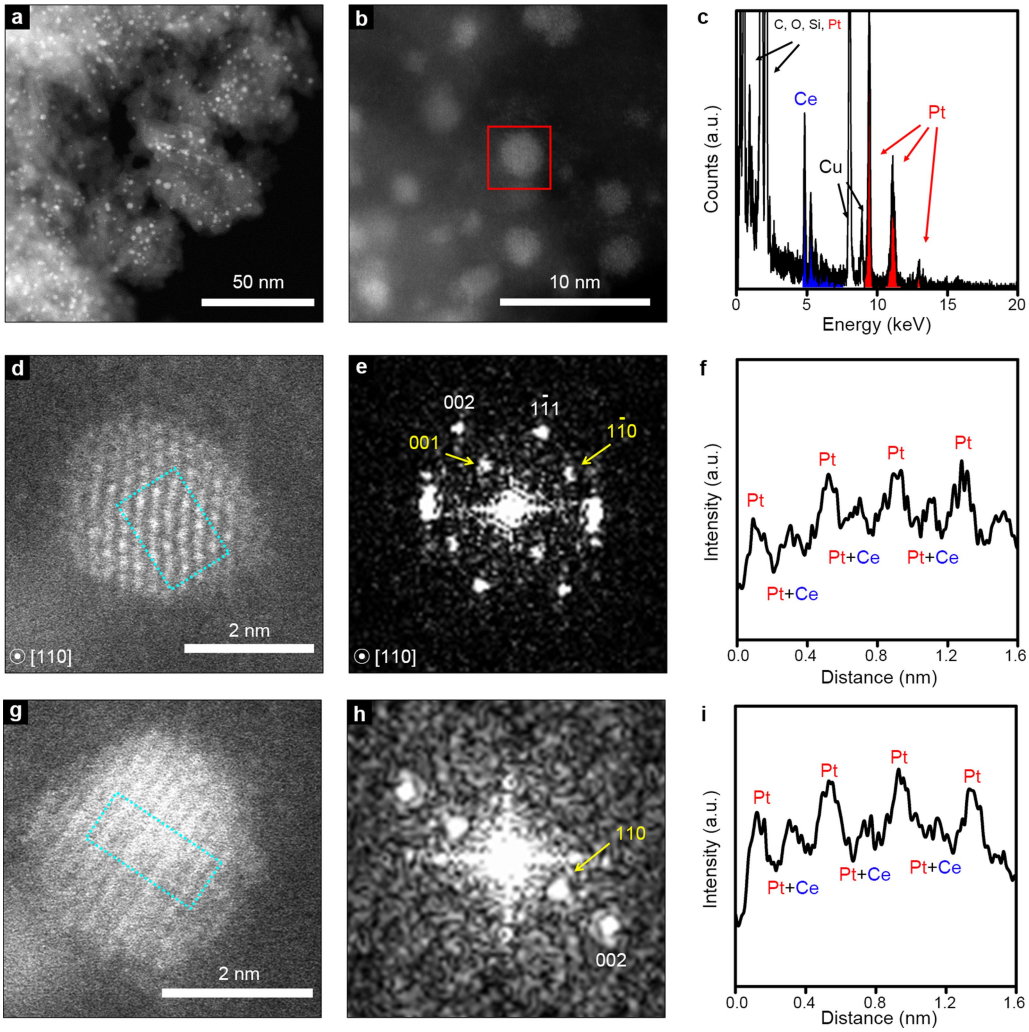
Sample	S_{BET} [a] ($cm^2 g^{-1}$)	V_{total} [b] ($cm^3 g^{-1}$)	d_{meso} [c] (nm)	Si/Ga[d]
Mesoporous MFI gallosilicate	660	1.17	5.2	48
mz-deGa	640	1.19	4.9	598
mz	660	0.83	3.0	n.d.[e]
alumina	170	0.45	9.5	n.d.[e]

[a]S_{BET} is BET surface area. [b]V_{total} is the total pore volume. [c]d_{meso} is the mode value of pore size distribution of each sample. [d]molar Si/Ga calculated by ICP-AES results. [e]Not determined.

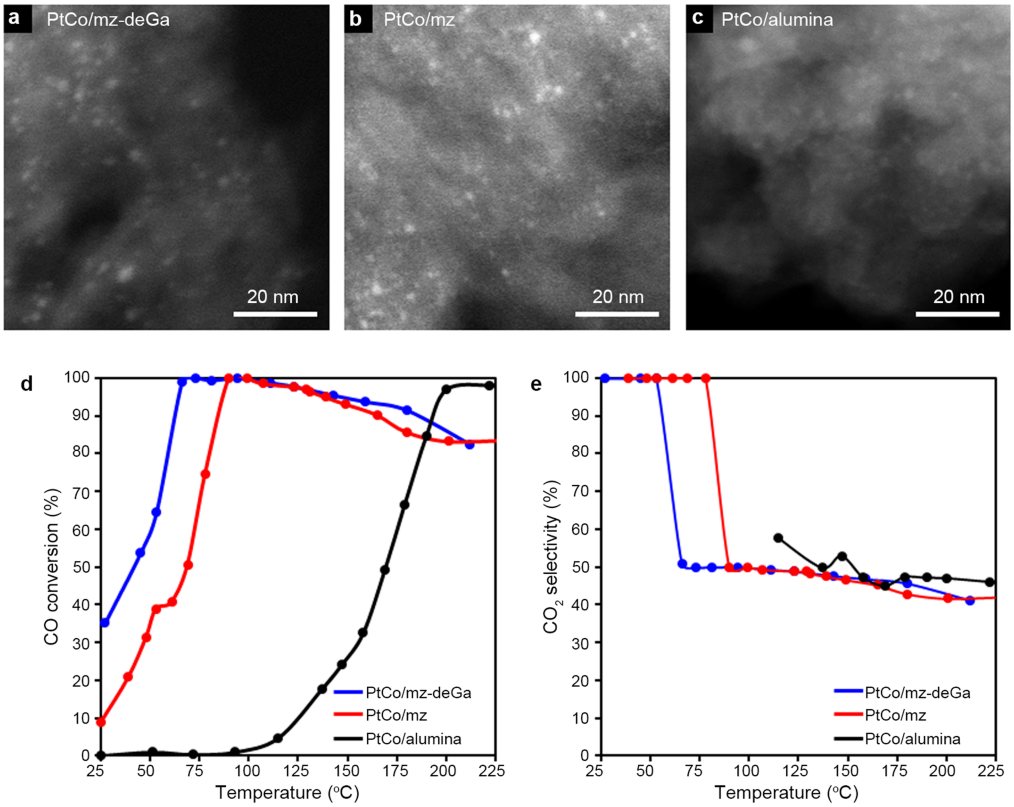
Extended Data Fig. 6 | Structure of mz siliceous mesoporous MFI synthesized using $C_{18\text{-}6\text{-}4}$ as a dual micro-/mesopore structure-directing agent. a, PXRD patterns. **b**, TEM image. **c**, N_2 adsorption–desorption isotherm. **d**, BJH pore size distribution derived from the adsorption branch of the N_2 isotherm. The mz zeolite showed an almost identical porous texture to that of the mesoporous MFI gallosilicate in Extended Data Fig. 2. **e**, Physicochemical properties of various zeolite samples and alumina.



Extended Data Fig. 7 | Propane dehydrogenation performance of catalysts using mz-deGa support. a, Propane conversion as a function of time on stream. **b**, Propylene selectivity as a function of time on stream. The supported catalysts are composed of either 1 wt% Pt or 1 wt% La. The PDH reaction conditions were as follows: 50 mg catalyst, WHSV = 11 h^{-1} with pure propane gas flow, and temperature of 580 °C. To determine the effect of remaining Ga species in mz-deGa on PDH performance, Pt/mz-deGa, La/mz-deGa and mz-deGa catalysts were tested. The mz-deGa catalyst showed negligible propane conversion, indicating that the remaining Ga species in the zeolite framework were not effective for PDH. Moreover, the La/mz-deGa sample showed negligible propane conversion, implying that the single atomic La species do not have PDH activity. Both mz-deGa and La/mz-deGa showed much poorer propylene selectivity compared to the Pt–REE/mz-deGa samples in Fig. 2. In the case of the Pt/mz-deGa sample, the initial propane conversion is slightly higher than that of Pt/mz shown in Fig. 2. In addition, the Pt/mz-deGa gave a noticeably lower deactivation rate than the Pt/mz sample. However, the improvement of Pt/mz-deGa in catalytic lifetime was almost insignificant when compared to that of the Pt–REE/mz-deGa catalysts. On the basis of these catalytic results, we can conclude that the remaining Ga species in mz-deGa could somehow promote the PDH performance of the supported Pt catalyst. However, this Ga contribution would be almost insignificant compared to that of Pt–REE alloy formation in the case of the Pt–REE/mz-deGa catalysts.



Extended Data Fig. 8 | Pt_3Ce nanoparticles with an $L1_2$ superlattice structure supported on mz-deGa. a, b, Low-magnification HAADF-STEM images showing uniformly sized metal nanoparticles dispersed on the zeolite. **c**, EDS spectrum taken from the red box in **b**, indicating the presence of both Pt and Ce in the same particle. **d, g**, AR-HAADF-STEM image of the metal nanoparticle, showing the Pt_3Ce ordered alloy structure with the $L1_2$ superlattice. **e**, **h**, FFT images from the HAADF-STEM image of **d** and **g**, respectively. The $L1_2$ superlattice reflection from the intermetallic compound structure is indicated by the yellow arrows in **e** and **h**. **f, i**, Intensity profiles taken along the light-blue boxes in **d** and **g**, respectively. The intensity profiles in **f** and **i** show that Pt and Pt + Ce atomic columns are alternating.

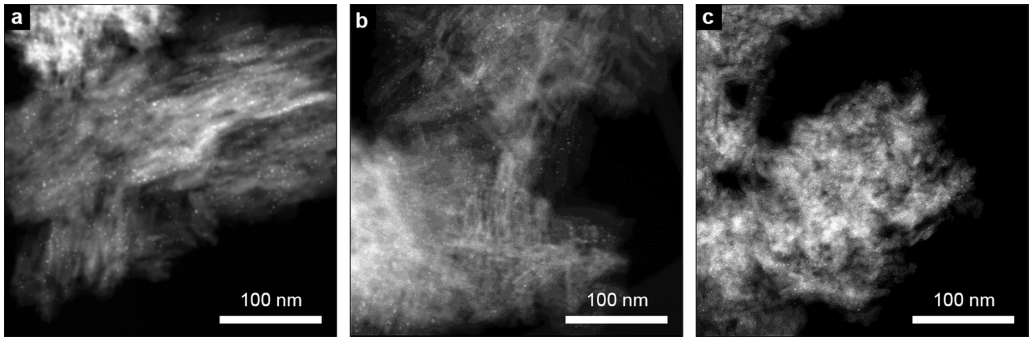


f

Sample	PtCo/mz-deGa	PtCo/mz	PtCo/alumina
Co/Pt mole ratio[a]	0.50	0.45	0.32

[a]Evaluated by selected-area STEM-EDS measurement

Extended Data Fig. 9 | Representative STEM images and preferential CO oxidation tests in a H_2-rich stream over supported PtCo catalysts. **a**–**c**, STEM images of PtCo/mz-deGa (**a**), PtCo/mz (**b**) and PtCo/alumina (**c**). **d**, **e**, CO conversion (**d**) and CO_2 selectivity (**e**) for supported PtCo catalysts. **f**, Amounts of Co species incorporated in Pt alloy nanoparticles supported on various supports measured with STEM-EDS. Depending on the support, there was a remarkable difference in the amounts of reduced Co species incorporated into the Pt alloy nanoparticles. As shown in **f**, the average Co/Pt molar ratio increased in the order PtCo/alumina (0.32) ≪ PtCo/mz (0.45) < PtCo/mz-deGa (0.50). The distinctively low Co/Pt of PtCo/alumina appears to be the result of the stronger interaction between the cationic Co species and the alumina support, which makes the reduction of Co more difficult than when using siliceous zeolite supports. Of the two siliceous zeolites, mz-deGa exhibited a larger degree of Co incorporation into the Pt nanoparticles. We believe that the silanol nests in mz-deGa could be beneficial to homogeneously distribute the Co species into atomic cations on the zeolite, thereby providing mobility to help atomistic diffusion. Such a benefit might enable better Co incorporation in the case of mz-deGa than in mz with no silanol nests. This difference in Co incorporation led to a variation in the catalytic performance for PROX in a H_2-rich stream containing 1.4% CO. The PtCo/mz-deGa catalyst exhibited complete CO conversion over a wide range of operating temperatures (60–120 °C). On the other hand, for the PtCo/mz catalyst complete CO conversion was achieved only above 90 °C. In the case of the PtCo/alumina catalyst, complete CO removal was not achievable, even at high temperatures.



Extended Data Fig. 10 | HAADF-STEM images. a, Pt/mz. **b**, PtLa/mz. **c**, PtSn/alumina. All three samples had uniformly dispersed Pt nanoparticles without any bulky sintered particles. These catalysts were tested for the PDH reaction, as shown in Fig. 2. For the preparation of Pt/mz, $Pt(NH_3)_4(NO_3)_2$ was loaded to mz using the incipient wetness impregnation technique. The Pt precursor-loaded mz was heated under an O_2 flow at 350 °C for 2 h (ramping rate, 0.8 °C min^{-1}; flow rate, 500 cm^3 min^{-1} g_{cat}^{-1}) and subsequently treated under a H_2 flow at 580 °C for 2 h (ramping rate, 0.3 °C min^{-1}; H_2 flow rate, 300 cm^3 min^{-1} g_{cat}^{-1}). The obtained Pt/mz contained 1 wt% Pt. For the preparation of PtSn/alumina, the alumina support was purchased from Sasol (PURALOX γ-Al_2O_3, 98%, BET surface area of 170 m^2 g^{-1}). The alumina was prepared by the same procedure as the Pt/mz except for the co-incorporation of H_2PtCl_6 and $SnCl_2{\cdot}2H_2O$ as metal precursors. The PtSn/alumina contained 1 wt% Pt and 1 wt% Sn.

Extended Data Table 1 | Hydrogen chemisorption results on various supported metal catalysts

Sample	Pt (wt%)	La (wt%)	Y (wt%)	H/Pt_{ini} [a]	H/Pt_{2nd} [b]	$(H/Pt_{2nd})/(H/Pt_{ini})$
Pt/mz	1			1.01	0.36	0.36
PtLa/mz-deGa	1	1		0.50	0.50	1.00
PtLa/mz	1	1		0.86	0.32	0.37
PtLa/alumina	1	1		0.48	0.22	0.46
PtY/mz-deGa	1		1	0.47	0.45	0.96
PtY/mz	1		1	0.83	0.25	0.30
PtY/alumina	1		1	0.57	0.22	0.39

This table shows hydrogen sorption results obtained at room temperature for various Pt–REE-supported zeolite and alumina samples. Except for PtLa/deGa and PtY/deGa, all the samples listed in the table showed that a substantial portion (60–70%) of initially adsorbed hydrogen was not desorbed upon evacuation for 1 h at room temperature, thereby resulting in considerably lowered H/Pt_{2nd} values compared to those of H/Pt_{ini}. The irreversible hydrogen sorption could be explained by the formation of a strong chemisorption bond of Pt–H. By contrast, the PtLa/mz-deGa and PtY/mz-deGa samples showed totally reversible hydrogen chemisorption behaviour, where the initial hydrogen uptakes were identical to those of the second hydrogen uptakes, which were taken after evacuation of the initially hydrogen sorption-measured samples for 1 h at room temperature. The completely reversible hydrogen sorption behaviours could indicate that the PtLa/mz-deGa and PtY/mz-deGa samples contained all the Pt nanoparticles as intermetallic compounds of Pt_3La and Pt_3Y. Therefore, all of the Pt portions in PtLa/mz-deGa and PtY/mz-deGa appeared to be alloyed with La or Y metal, forming Pt_3La or Pt_3Y intermetallic compounds. Considering this, ~70 mol% of La and ~80 mol% of Y would exist in PtLa/mz-deGa and PtY/mz-deGa, respectively, as isolated from Pt. The remaining REEs appeared as single atomic species on the zeolite supports.

[a]Molar H/Pt value at the initial chemisorption measurement at room temperature.

[b]Molar H/Pt value at the second chemisorption measurement after hydrogen desorption by evacuation for 1 h.

Satellite isoprene retrievals constrain emissions and atmospheric oxidation

https://doi.org/10.1038/s41586-020-2664-3

Received: 17 August 2019

Accepted: 14 July 2020

Published online: 9 September 2020

Check for updates

Kelley C. Wells[1], Dylan B. Millet[1]✉, Vivienne H. Payne[2], M. Julian Deventer[1,11], Kelvin H. Bates[3], Joost A. de Gouw[4,5], Martin Graus[6], Carsten Warneke[4,7], Armin Wisthaler[8,9] & Jose D. Fuentes[10]

Isoprene is the dominant non-methane organic compound emitted to the atmosphere[1–3]. It drives ozone and aerosol production, modulates atmospheric oxidation and interacts with the global nitrogen cycle[4–8]. Isoprene emissions are highly uncertain[1,9], as is the nonlinear chemistry coupling isoprene and the hydroxyl radical, OH—its primary sink[10–13]. Here we present global isoprene measurements taken from space using the Cross-track Infrared Sounder. Together with observations of formaldehyde, an isoprene oxidation product, these measurements provide constraints on isoprene emissions and atmospheric oxidation. We find that the isoprene–formaldehyde relationships measured from space are broadly consistent with the current understanding of isoprene–OH chemistry, with no indication of missing OH recycling at low nitrogen oxide concentrations. We analyse these datasets over four global isoprene hotspots in relation to model predictions, and present a quantification of isoprene emissions based directly on satellite measurements of isoprene itself. A major discrepancy emerges over Amazonia, where current underestimates of natural nitrogen oxide emissions bias modelled OH and hence isoprene. Over southern Africa, we find that a prominent isoprene hotspot is missing from bottom-up predictions. A multi-year analysis sheds light on interannual isoprene variability, and suggests the influence of the El Niño/Southern Oscillation.

Isoprene (2-methyl-1,3-butadiene), produced during photosynthetic metabolism and emitted mainly from the leaves of woody plants, has global emissions comparable to those of methane and considerably greater than the sum of anthropogenic volatile organic compounds (VOCs)[1–3]. Isoprene is highly reactive (lifetime <1 h at [OH] = 5×10^6 molecules per cm^3) and plays a pivotal role in atmospheric oxidation, ozone and aerosol formation[4–8]. Air quality and chemistry–climate models thus require accurate isoprene emission inputs; however, current estimates span a wide range (~210–990 Tg C yr^{-1} globally[1,9]). The degree to which isoprene oxidation at low nitrogen oxide (NO_x) levels depletes versus sustains the abundance of hydroxyl radicals (OH)—the principal atmospheric oxidant[10–13]—is also uncertain. Space-borne measurements of formaldehyde (HCHO, an isoprene oxidation product) can provide top-down constraints[14], but alone its use as an isoprene proxy is hampered by uncertainties in the NO_x-dependent chemistry governing the formaldehyde production yield and timescale[15], and by competing non-isoprene formaldehyde sources[14,16–18].

Fu et al.[19] recently demonstrated the viability of direct space-borne isoprene retrievals using infrared (IR) radiance measurements from the Cross-track Infrared Sounder (CrIS). That study employed optimal estimation to retrieve isoprene column abundances ($\Omega_{isoprene}$; see Supplementary Note 1 and Supplementary Fig. 1) over Amazonia, with results validated using aircraft measurements. Here, we build on that work to develop an artificial neural network (ANN)-based algorithm for deriving global isoprene columns from the CrIS measurements. The computational efficiency of the ANN allows fuller exploitation of the dense CrIS sampling (~9×10^6 spectra per day) for understanding spatial and temporal drivers of atmospheric isoprene. We thus derive global observations of atmospheric isoprene from space, and use this dataset to evaluate current understanding of its emissions and atmospheric oxidation.

Isoprene spectral index

As described in Methods, we use the CrIS-measured brightness temperature difference (ΔT_b) between the peak of the ν_{28} isoprene band[20] and nearby off-peak channels (see Extended Data Fig. 1a) as a spectral index for deriving isoprene column abundances from the satellite data. Analogous methodologies have been used successfully for a variety of other atmospheric species[21–24]. Extended Data Fig. 1b shows the ΔT_b–isoprene relationship as simulated by a forward radiative transfer model for diverse conditions spanning the global atmosphere over land (Methods). The relationship is approximately linear with slope varying as a function of thermal contrast (atmosphere–surface temperature difference; see Methods and Extended Data Fig. 1c). Interfering species likewise play a role and need to be accounted for, as discussed later.

[1]Department of Soil, Water, and Climate, University of Minnesota, St Paul, MN, USA. [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [4]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [5]Department of Chemistry, University of Colorado, Boulder, CO, USA. [6]Department of Atmospheric and Cryogenic Sciences, University of Innsbruck, Innsbruck, Austria. [7]NOAA Chemical Sciences Laboratory, Boulder, CO, USA. [8]Institute for Ion Physics and Applied Physics, University of Innsbruck, Innsbruck, Austria. [9]Department of Chemistry, University of Oslo, Oslo, Norway. [10]Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, PA, USA. [11]Present address: Bioclimatology, University of Göttingen, Göttingen, Germany. ✉e-mail: dbm@umn.edu

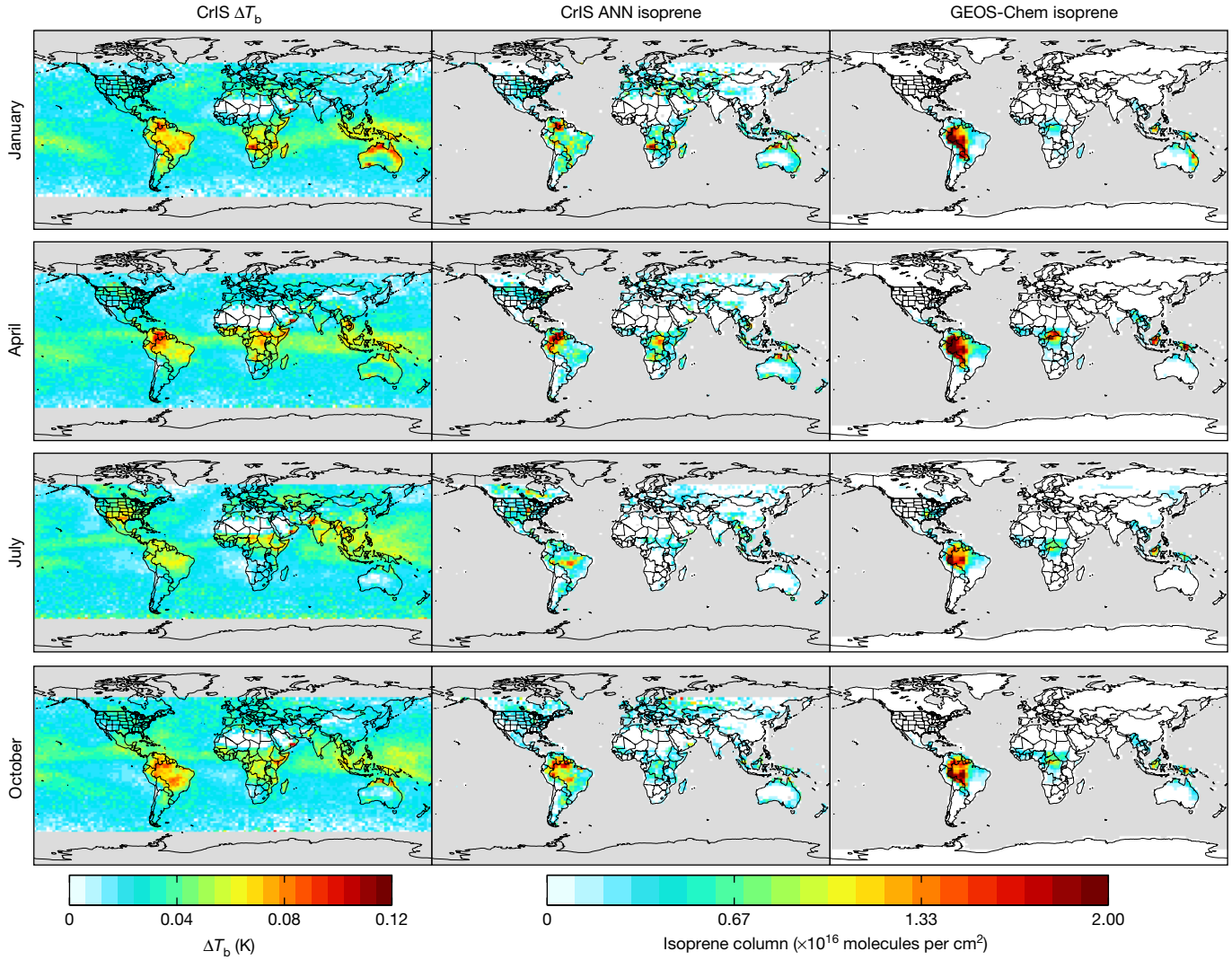


Fig. 1 | Global distribution of ΔT_b and isoprene columns. Left, monthly mean ΔT_b observations from CrIS. Middle, isoprene column densities derived from the CrIS observations. Right, isoprene column densities simulated by GEOS-Chem. Data are plotted for January, April, July and October 2013 at ~13:30 LT (12:00–15:00 LT mean, with daily cloud screening applied; LT, local time). Ocean and high-latitude pixels (in grey) are excluded from the isoprene maps as they are not part of the ANN training dataset (see Methods).

Figure 1 maps the global and seasonal ΔT_b distribution measured by CrIS. Clear enhancements are seen over many predicted isoprene source regions: Amazonia, northern Australia (January), central Africa (April) and the southeast United States (July). However, ΔT_b enhancements also manifest over regions not predicted by the GEOS-Chem chemical transport model (CTM; Methods) to have large isoprene sources (for example, equatorial eastern Africa and the Arabian Peninsula, Pakistan and the southwest United States in July, Angola/Zambia in January and April). Elevated ΔT_b values also occur across the tropics, with a spatial distribution resembling that of water vapour. As will be seen, ΔT_b enhancements not associated with high modelled isoprene can reveal locations where emissions are much higher than presently thought—many parts of the world lack flux measurements for regionally important plant species. However, we show later that the rest of these anomalous features disappear once thermal contrast, water vapour and related factors are properly accounted for via the ANN.

ANN-based isoprene measurements

We use a supervised feed-forward ANN to derive isoprene columns from the CrIS ΔT_b data and contemporaneous observations of relevant surface and atmospheric properties (Methods). The ANN used (representing the mean of 10 networks) reproduces 93% of the isoprene column variance across the full training dataset. Prediction uncertainty is typically <30% for elevated isoprene columns (>1×10^{16} molecules per cm^2), increasing to 50% or more for low isoprene amounts/low thermal contrast.

We apply the trained ANN to the space-borne CrIS ΔT_b measurements to derive global isoprene distributions for January, April, July and October 2013 (Methods). Because the statistical performance of the ANN summarized above does not necessarily represent the full observational uncertainty, we further evaluate our results with the previously validated optimal estimation retrievals[19] and with independent aircraft measurements from two campaigns over the southeast United States[25,26].

Figure 2a compares the ANN and optimal estimation isoprene measurements over Amazonia for September 2014, revealing strong agreement between the two (correlation $r = 0.9$, slope $m = 0.8$). Furthermore, Fig. 2b–d shows that the aircraft–model comparisons (see Methods) yield slopes ($m = 1.2$–1.3) and correlations ($r = 0.5$–0.7) that statistically match the CrIS–model comparison ($m = 1.3$, $r = 0.6$), thus providing indirect validation of the CrIS data. The aircraft measurements also reveal key spatial features that are consistent with CrIS but not captured by

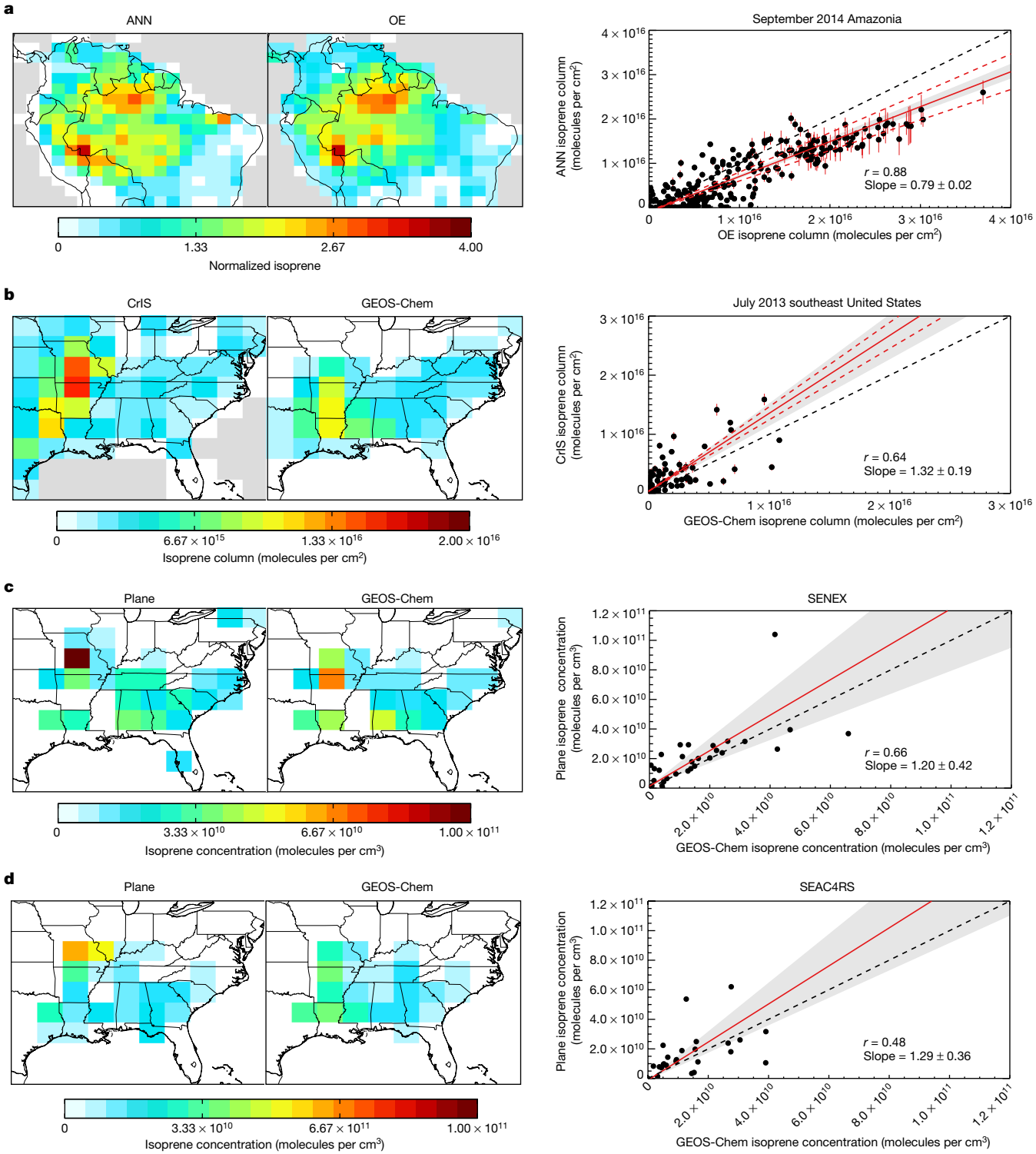


Fig. 2 | Comparison of the CrIS ANN isoprene columns with other datasets. **a**, Comparison of ANN- and optimal estimation (OE[19])-derived isoprene estimates. Both are derived from cloud-screened CrIS radiance data for September 2014; ANN results employ GEOS-Chem HNO_3 as CrIS HNO_3 data were unavailable for this timeframe. The maps (left) display columns normalized to their domain means, and the scatterplot (right) compares the absolute columns (absolute columns are mapped in Extended Data Fig. 8). **b**–**d**, Evaluation of CrIS ANN isoprene measurements using aircraft observations and GEOS-Chem model output. **b**, Monthly mean July 2013 isoprene columns as measured by CrIS (~13:30 LT) and simulated by GEOS-Chem (12:00–15:00 LT mean). **c**, **d**, Ambient isoprene concentrations as measured during the SENEX (June–July 2013; **c**) and SEAC[4]RS (August–September 2013; **d**) aircraft campaigns and simulated by GEOS-Chem along the flight tracks. Data are plotted as campaign-average density-weighted boundary layer number densities ($P > 800$ hPa). The error bars in the scatter plots for **a** and **b** indicate the standard deviation across the 10 ANN-based columns (see Methods; in some cases error bars are smaller than the data points), the red dashed lines indicate the range in slopes across ANNs and black dashed lines indicate the 1:1 relation. Stated slope uncertainties and grey shaded regions represent the bootstrapped standard error of regression.

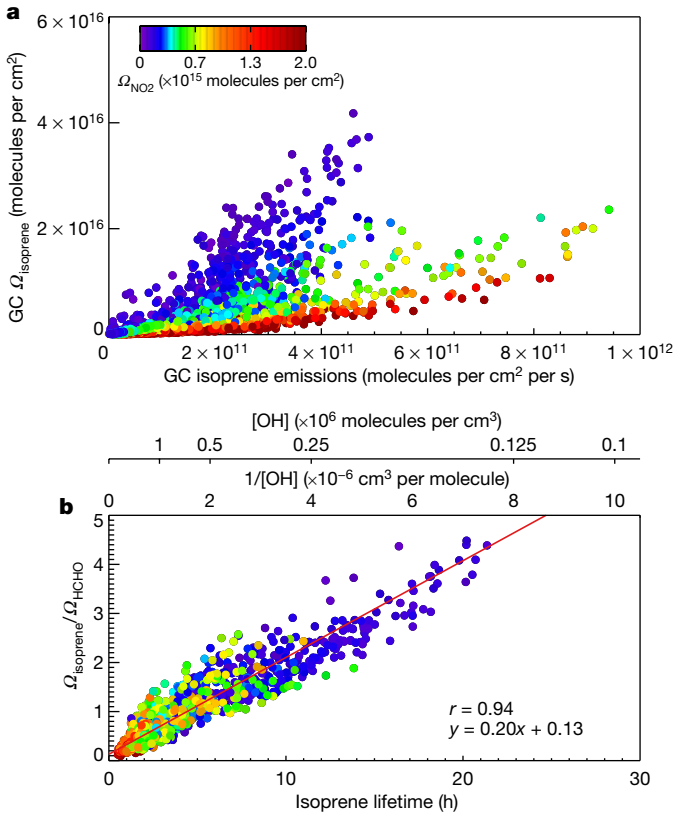


Fig. 3 | Dependence of atmospheric isoprene columns on emissions and lifetime. a, The global ensemble of monthly-mean ~13:30 LT (12:00–15:00 LT mean) GEOS-Chem (GC) isoprene columns predicted for 2013 versus the corresponding isoprene emissions. **b**, The predicted isoprene:HCHO column ratio shown as a function of isoprene lifetime, 1/[OH] and [OH] (all for z < 500 m). The data points in both plots are colour coded by the modelled tropospheric NO_2 column.

GEOS-Chem. In particular, the largest observed isoprene enhancements occur over the Missouri Ozarks, farther north than model predictions—as also seen by CrIS. Finally, the enhancement magnitude measured during both aircraft campaigns is larger than predicted by the model, a finding likewise reflected in the CrIS data (Fig. 2b–d).

These results provide robust support for the ANN-derived isoprene abundances from CrIS. Looking forward, more validation datasets in high-isoprene regions (specifically, airborne or surface-based column measurements) would enable more extensive uncertainty assessment and retrieval improvement.

CrIS isoprene reflects emissions and OH

The global isoprene column distribution is governed by the balance between emissions and loss (predominantly via reaction with OH). Extended Data Figure 2 maps global isoprene emissions, lifetimes and columns predicted by GEOS-Chem. Because modelled OH (and therefore the isoprene lifetime) varies strongly with NO_x and with isoprene itself, the isoprene distribution differs substantially from that of emissions. For example, predicted July emissions are higher in the southeast United States than Amazonia, yet the resulting isoprene columns are dramatically higher over Amazonia.

Figure 3a quantifies this effect in the model by plotting the global ensemble of monthly mean 13:30 LT isoprene columns against emissions. Points are colour coded by simulated tropospheric nitrogen dioxide (NO_2) columns, and two limiting regimes emerge. At elevated NO_x ($\Omega_{NO2} \gtrsim 10^{15}$ molecules per cm^2) the relationship is near linear, reflecting an approximate local steady state between isoprene columns and emissions, with the slope corresponding to the isoprene lifetime. At lower NO_x, the isoprene columns increase superlinearly with emissions. In this regime (occurring in the model most notably over Amazonia), elevated isoprene suppresses OH and therefore its own sink, leading to runaway concentrations.

As an isoprene oxidation product, formaldehyde is more buffered with respect to OH variability: (1) photolysis ensures that HCHO removal continues even at low OH levels, and (2) its production is proportional to isoprene × OH, which is more stable than either quantity alone when elevated isoprene suppresses OH. Because of these differing sensitivities, the isoprene:HCHO column ratio is a proxy for the atmosphere's oxidizing capacity over isoprene source regions. Figure 3b illustrates this relationship: on a global basis, across all locations and seasons, the monthly mean 13:30 LT $\Omega_{isoprene}/\Omega_{HCHO}$ ratios simulated by GEOS-Chem scale tightly with 1/[OH] (r = 0.94; Supplementary Note 2 discusses the factors driving this relationship). A sensitivity analysis using an alternate isoprene oxidation mechanism (Mini-CIM[8]; see Methods) yields a similarly strong correlation (Supplementary Fig. 2), with details presented in Supplementary Note 3.

The strong correlation in Fig. 3b encompasses the full global range of chemical regimes for isoprene oxidation: from unpolluted situations where isoprene-derived peroxy radicals (RO_2) are long-lived and react mainly with hydroperoxyl radicals (HO_2) and other RO_2 or isomerize, to polluted areas where isoprene-derived RO_2 react quickly with NO (refs. [27,28]). This globally aggregated $\Omega_{isoprene}/\Omega_{HCHO}$ versus 1/[OH] slope is weighted to isoprene-rich, OH-poor conditions: Supplementary Note 3 shows that the modelled slope varies across our analysis regions from 0.18 to 0.49. A sensitivity study with the independent Mini-CIM mechanism further shows systematic adjustments of 28–56% depending on location (Supplementary Fig. 4); factors such as non-isoprene biogenic VOC emissions and model mixing assumptions (which influence the column-integrated OH–isoprene reaction rate[29]) also influence the slope (Supplementary Note 3). Overall, however, results here clearly demonstrate that the $\Omega_{isoprene}/\Omega_{HCHO}$ ratio provides a strong proxy of atmospheric oxidation that is observable from space.

We can therefore derive new constraints on isoprene–OH chemistry globally by combining the CrIS isoprene measurements derived here with space-based HCHO columns from OMI[30] (Ozone Monitoring Instrument; Methods). Specifically, we employ the measured isoprene:HCHO ratios from CrIS and OMI as a direct proxy of 1/[OH] (and hence the isoprene lifetime) that can be used to test chemical models. To that end, Fig. 4 plots the $\Omega_{isoprene}/\Omega_{HCHO}$ ratios measured by CrIS + OMI and simulated by GEOS-Chem. Data are shown as a function of isoprene and NO_2 (from ref. [31]) for months spanning all four seasons (January, April, July and October), and confined to locations with elevated surface temperatures (>293 K) to limit noise due to low isoprene/thermal contrast. In both satellite-based and modelled relationships, we see a low-OH (and long isoprene lifetime) regime when isoprene is elevated and NO_x is low, and an opposing higher-OH (short lifetime) regime when the reverse is true. These oxidative regimes, and the chemical transitions between them, are generally consistent between model and observations, with the corresponding $\Omega_{isoprene}/\Omega_{HCHO}$ ratios (and thus OH) agreeing to within 10–40% at low to moderate NO_2 ($\lesssim 10^{15}$ molecules per cm^2). One clear discrepancy is that the model population of extremely high isoprene at extremely low NO_x is not seen in the data; as we will see, this primarily reflects model NO_x errors over Amazonia. Some disparities also emerge at elevated NO_2; however, the observed values in this range are subject to greater error due to limited measurements and lower isoprene columns with more uncertainty (Extended Data Fig. 3).

The above comparison supports the current model treatment of OH chemistry in the presence of isoprene. In particular, it argues against any substantial missing OH recycling at low NO_x (refs. [10,12,32])—instead, the modelled OH levels are modestly higher than implied by the satellite data. A sensitivity analysis using the Mini-CIM[8] isoprene oxidation

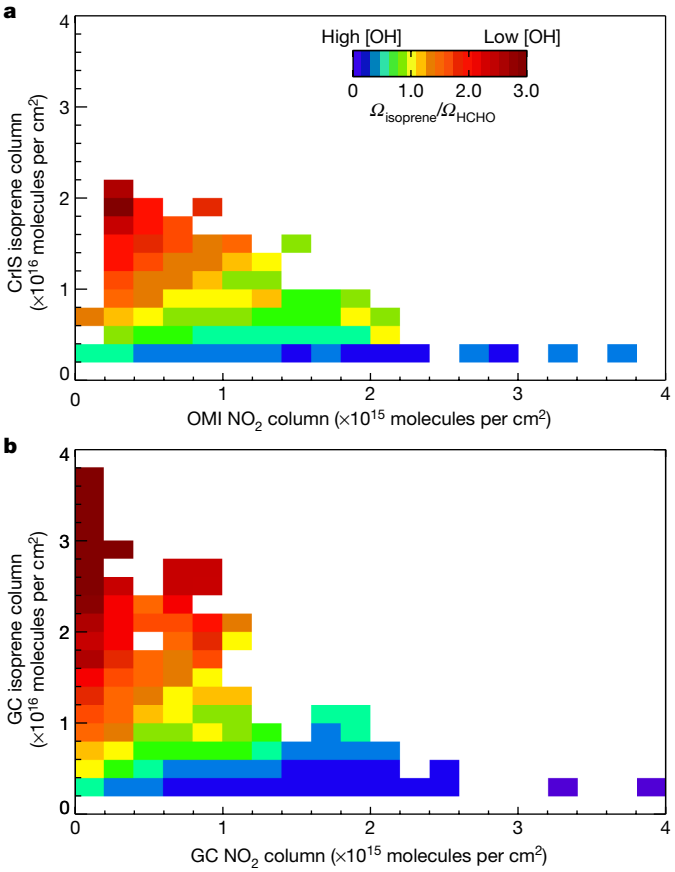


Fig. 4 | Global distribution of the isoprene:HCHO ratio as a function of isoprene and NO_x. The isoprene:HCHO ratio is a proxy for 1/[OH] (Fig. 3). **a**, The observed relationship based on CrIS and OMI data. **b**, The simulated relationship from GEOS-Chem. Plotted ratios represent monthly mean values at 13:30 LT (12:00–15:00 LT mean) and are binned by isoprene and tropospheric NO_2 column amounts. Data reflect locations with elevated surface temperature (>293 K at satellite overpass) and where the isoprene and HCHO measurements are above detection limit (2×10^{15} molecules per cm^2).

mechanism supports this conclusion (Supplementary Note 4 and Supplementary Fig. 5). In the following, we therefore examine the CrIS isoprene distribution and seasonality in light of the oxidative information provided by the $\Omega_{isoprene}/\Omega_{HCHO}$ ratio, with measurement–model differences used to inform present understanding of emissions and atmospheric NO_x.

Figure 1 shows the global CrIS isoprene columns and corresponding GEOS-Chem predictions for January, April, July and October 2013. The CrIS data reveal a number of isoprene hotspots that are consistent with the known isoprene sources discussed earlier—in particular, Amazonia, Central Africa, Australia and the Ozarks of the southeast United States. These regions stand out because they combine strong emissions with a chemical regime where isoprene is sufficiently long-lived to be detectable from space (unlike, for example, China in July, with elevated emissions but shorter isoprene lifetimes; Extended Data Fig. 2). For the months shown, the Central Africa and southeast United States enhancements peak in April and July, respectively, consistent with model predictions.

These dominant isoprene features are robust across the suite of ANN predictions: the column standard deviation across networks is typically <25% in these regions (Methods; Extended Data Fig. 4). The anomalous ΔT_b enhancements discussed earlier in the context of spectral interferences do not emerge as enhancements in the CrIS isoprene maps, showing that the ANN is effectively accounting for non-isoprene factors influencing ΔT_b. A notable feature not predicted by GEOS-Chem is the strong observed isoprene enhancement over southern Africa in January and, to a lesser degree, in April; this is explored later in the text.

The following sections examine each of the above hotspots in terms of their implications for present understanding of atmospheric isoprene. For each region, we apply the corresponding $\Omega_{isoprene}/\Omega_{HCHO}$ versus 1/[OH] relationship in Supplementary Fig. 3 as a transfer function to quantify OH and the isoprene lifetime from the measured isoprene:HCHO ratios. The same transfer function is likewise applied to the model ratios (in this way, all relative model–measurement lifetime discrepancies arise solely from the underlying isoprene and HCHO column data, and are unaffected by any transfer function uncertainty). We also use the satellite measurements to provide an initial quantification of isoprene (and NO_x) emissions over the same global hotspots, as detailed in Supplementary Notes 5, 6. From this analysis, we identify and discuss emergent gaps in current bottom-up understanding of isoprene emissions. Results are summarized in Figs. 5, 6 and in Supplementary Figs. 6, 7, 12–17.

Amazonia. The CrIS isoprene columns over Amazonia reveal strong seasonal variability in both the magnitude and location of the isoprene maxima. For the months examined, observed columns in west Amazonia (Fig. 5b, Extended Data Fig. 5) are highest in October and April and lowest in July. This is consistent with local ground-based measurements during GoAmazon[33], which exhibit a June minimum and increase nearly twofold from then to October (Fig. 5b). Wei et al.[33] attribute this seasonal minimum to leaf-flushing between wet and dry seasons; other studies[34,35] also infer low isoprene emissions during new leaf growth in June–July. This seasonality is not well represented in GEOS-Chem, which instead peaks in April and exhibits only a 5% July–October column increase.

Also apparent from Fig. 5a is that the regions with long isoprene lifetimes and low OH concentrations based on the $\Omega_{isoprene}/\Omega_{HCHO}$ observations also have low NO_x concentrations based on OMI NO_2 (for example, $\Omega_{NO2} < 0.2\times10^{15}$ molecules per cm^2 corresponds here in GEOS-Chem to surface [NO] < 32 parts per trillion (ppt) and RO_2 lifetimes to NO of >2.4 min), especially in January, April and October. This agrees with chemical expectations for isoprene-rich, NO_x-poor environments and thus provides strong confirmation of our approach, since the lifetime/OH constraints are derived only from isoprene and HCHO without incorporating any NO_x data.

Whereas the measured isoprene columns reveal localized maxima varying by season, GEOS-Chem instead predicts persistently elevated isoprene throughout much of western Amazonia. The model simultaneously predicts a much broader region of low OH (and elevated isoprene lifetime) than is inferred from the satellite data from January–July. We attribute these discrepancies mainly to the marked, widespread model NO_x underestimation apparent during these months (Fig. 5a). Although the modelled Amazonian NO_x levels are frequently low enough to yield the runaway isoprene concentrations discussed previously, the observations do not show this occurring to such an extent.

Simulations using the Mini-CIM mechanism (Supplementary Fig. 6), despite featuring some spatial differences compared with the standard model, nonetheless lead to similar overall conclusions when evaluated against the satellite data. Specifically, predicted isoprene columns from January–July are higher than is observed, although to a lesser degree than in the base-case simulations due to higher OH values in Mini-CIM. Furthermore, the suppressed OH levels predicted by Mini-CIM over Amazonia extend over a broader geographic area than is revealed in the satellite data. As before, this disparity exhibits a spatial fingerprint that matches the overly low model NO_x values, as implied by OMI Ω_{NO2}.

The above Ω_{NO2} bias could theoretically reflect model NO_x errors in the free troposphere or boundary layer[36,37]; the former would have little effect on near-surface isoprene chemistry. However, we find that GEOS-Chem surface NO_x predictions are indeed too low relative to

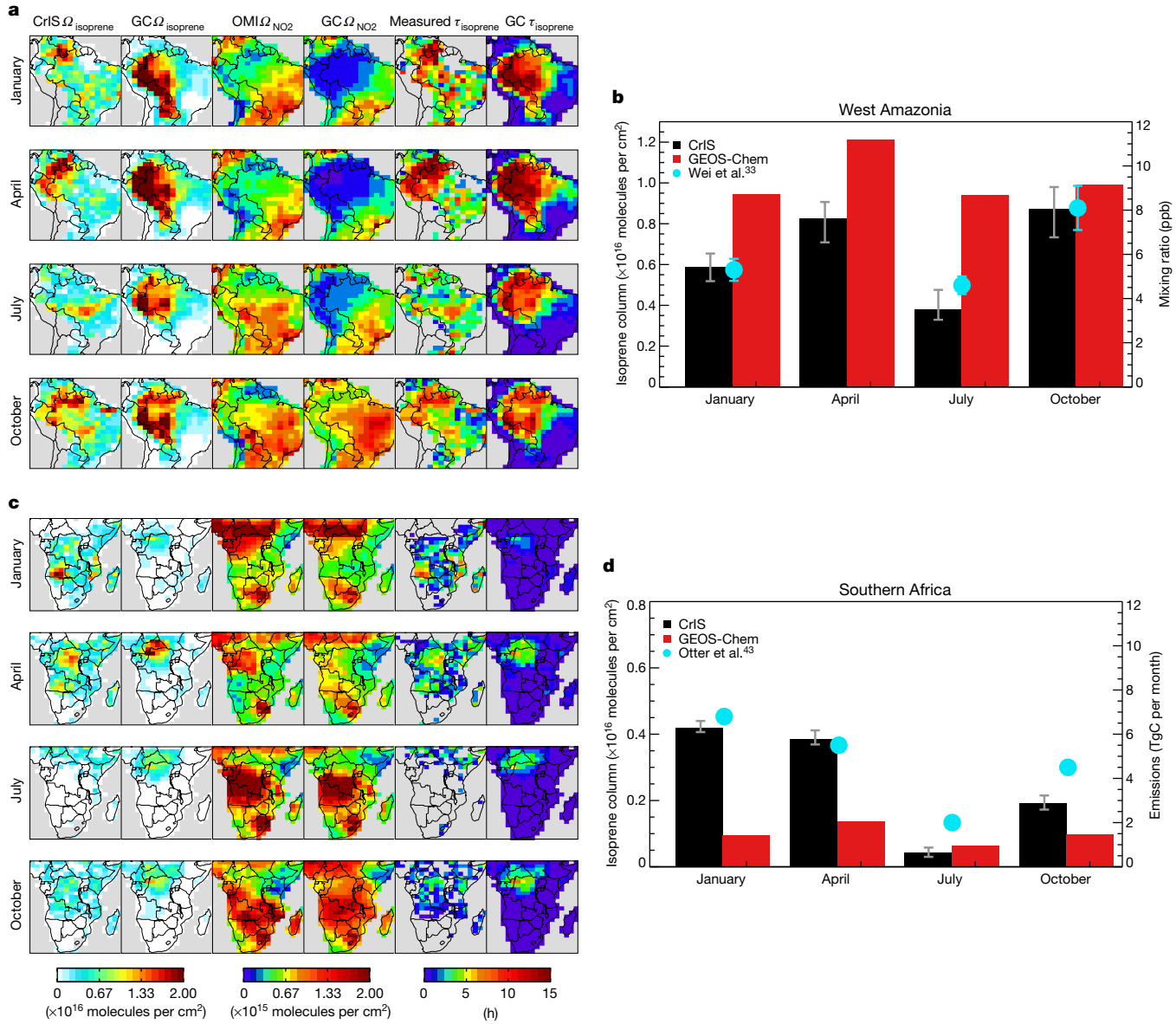


Fig. 5 | Seasonality of space-based isoprene over Amazonia and southern Africa. a, **c**, CrIS and GEOS-Chem isoprene columns, OMI and GEOS-Chem tropospheric NO_2 columns and space-based and GEOS-Chem isoprene lifetimes ($\tau_{isoprene}$, calculated from the isoprene:HCHO ratios via the transfer functions of Supplementary Fig. 3) for Amazonia (**a**) and Southern Africa (**c**) during January, April, July and October 2013 (top to bottom). The CrIS isoprene and space-based isoprene lifetimes are shown for snow-free, above detection limit locations ($\Omega_{isoprene}$ and Ω_{HCHO} values >2×10^{15} molecules per cm^2). **b**, **d**, Regional mean CrIS (black; error bars indicate the range across ANN predictions) and GEOS-Chem (red) isoprene columns for west Amazonia (**b**; regions defined in Extended Data Fig. 5) and southern Africa (**d**). Results for west Amazonia are compared with in situ mixing ratios (cyan points and error bars show the 12:00–15:00 LT mean and standard deviation) measured from May 2014–January 2015 in the central Amazon Basin[33]. In situ data were unavailable for most of July, so the July CrIS values are compared with the in situ mean for June 2014. Results for Southern Africa are compared with monthly isoprene emissions from a detailed regional inventory[43].

surface observations during GoAmazon[38] (Supplementary Note 5). From in situ measurements, Liu et al.[39] likewise infer a large near-surface NO_x bias in GEOS-Chem predictions for this region, which they attribute to underestimated soil emissions. Our satellite-based optimization described in Supplementary Note 5 (Supplementary Figs. 9–11, Supplementary Table 1) leads to substantial Amazonian NO_x emission increases that agree well with the findings of Liu et al.[39].

Supplementary Figure 11 further shows that our NO_x optimization successfully reduces the large isoprene lifetime biases over Amazonia in the prior model—providing independent confirmation of the results and supporting this isoprene emission quantification using CrIS. We thus derive monthly Amazonian isoprene emissions that point to substantial and coherent spatial errors in the bottom-up inventory (details in Supplementary Note 5). Overall, these results highlight the critical need to better understand NO_x sources for this part of the world, and to elucidate the mechanisms driving isoprene emission variability in the tropics.

Africa. Two African isoprene hotspots are observed by CrIS: one in central Africa in April and one in the Miombo and transitional woodlands of Angola peaking in January (Fig. 5c–d)[40]. Although GEOS-Chem captures the timing of the central African enhancement, the CrIS data show the predicted isoprene peak to be too strong and too far north—as found previously based on OMI HCHO[41] (model predictions using Mini-CIM are similar; Supplementary Fig. 6).

The Miombo/Angola peak has not been previously identified to this extent, though elevated leaf-level isoprene fluxes have been observed

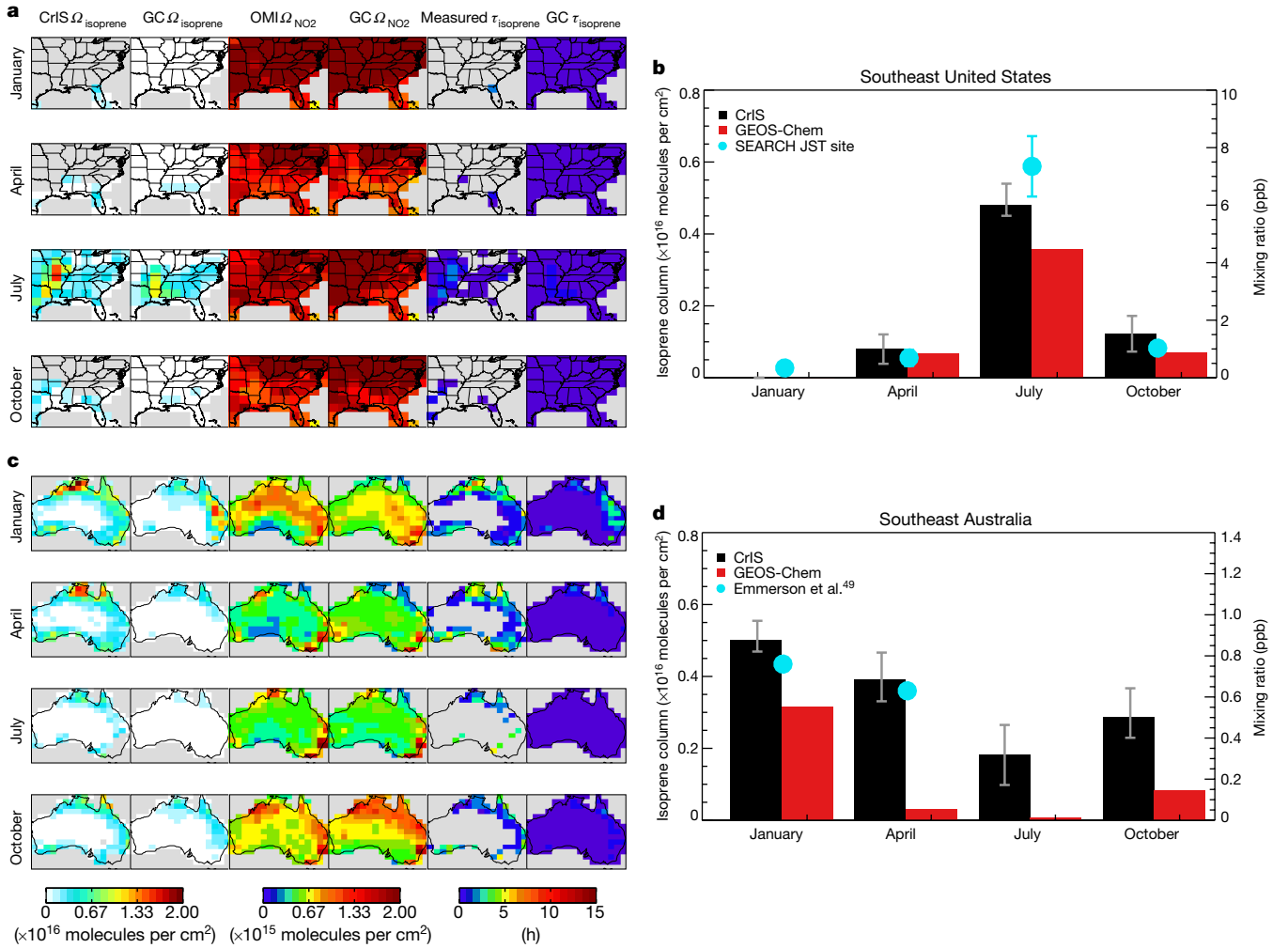


Fig. 6 | Seasonality of space-based isoprene over the southeast United States and Australia. a, **c**, CrIS and GEOS-Chem isoprene columns, OMI and GEOS-Chem tropospheric NO_2 columns and space-based and GEOS-Chem isoprene lifetimes for the southeast United States (**a**) and southeast Australia (**c**) during January, April, July and October 2013 (top to bottom). The CrIS isoprene and space-based isoprene lifetimes are shown for snow-free, above detection limit locations ($\Omega_{isoprene}$ and Ω_{HCHO} values >2×10^{15} molecules per cm^2). **b**, **d**, Regional mean CrIS (black; error bars indicate the range across ANN predictions) and GEOS-Chem (red) isoprene columns for the southeast United States (**b**; regions defined in Extended Data Fig. 5) and southeast Australia (**d**). The southeast United States results are compared with 10-yr mean (1999–2008) isoprene concentration measurements from Atlanta, GA (ref. [48]; cyan; error bars indicate the 10-yr standard deviation). Southeast Australian results are compared with measurements from the Sydney Particle Study[49]. January and April CrIS values are compared with summer (1 February–7 March 2011) and autumn (14 April–14 May 2012) campaign means, respectively.

in woody savannahs here[42]. Furthermore, while the CrIS-observed hotspot is largely missing from MEGANv2.1 (ref. [1]), it matches the location and season of highest emissions according to a regional inventory from Otter et al.[43] that incorporates detailed local land-cover information. The enhancement occurs in a low-NO_x (and therefore low OH) area, leading to large isoprene enhancements relative to the corresponding emissions and HCHO (Figs. 3a, 5c) and explaining why a correspondingly strong HCHO peak is not seen. The CrIS seasonality over southern Africa also compares well with Otter et al.[43] (Fig. 5d, Extended Data Fig. 5), with a January maximum and July minimum. GEOS-Chem, conversely, peaks in April with isoprene columns 2–4 times lower than CrIS.

The total isoprene emissions inferred from CrIS over southern Africa are higher than the prior estimate during January and April (Supplementary Fig. 13), and imply an emission overestimate north of the Equator and underestimate to the south (particularly over Angola/Namibia). These emission adjustments broadly support previous HCHO-based findings[14,41,44,45]. As described in Supplementary Note 5, our CrIS-derived isoprene emissions for all of sub-equatorial Africa are highly consistent with the Otter et al.[43] estimates, but substantially higher than MEGANv2.1. Such large discrepancies reveal a need for further investigation of isoprene sources in this understudied region.

Southeast United States. CrIS isoprene columns over the southeast United States peak in July over the 'isoprene volcano' in Missouri/Arkansas, where surface mixing ratios up to 36 parts per billion (ppb) have been observed[46]. The aircraft data shown in Fig. 2b–d corroborate the CrIS isoprene distribution over this region, and OMI HCHO columns (Extended Data Fig. 6) likewise peak over the same part of the Ozarks during this time.

The GEOS-Chem isoprene maximum is shifted southwards with lower column amounts than CrIS (Fig. 6a, Supplementary Fig. 7). Kaiser et al.[47] emphasize the importance of correcting NO_x biases when inferring isoprene emissions, and modelled NO_2 columns exhibit substantial, spatially varying biases over this region (Fig. 6a and Supplementary Fig. 8). The isoprene lifetime predicted by the standard model is ~2 times the satellite-inferred value over the southern portion of the domain (where model isoprene is biased high) and 30–50% too low over Missouri (where the model is too low). However, the model does capture the observed regional isoprene seasonality[48] (Fig. 6b). After

correcting the NO_x biases above, we derive from the CrIS data moderate downward isoprene emission adjustments over Louisiana, Mississippi and Alabama, offset by increases over Missouri, Illinois and eastern Texas (Supplementary Fig. 14).

Australia. CrIS isoprene columns over Australia are highest in the north during January and April, with smaller enhancements along portions of the eastern and southern coasts (Fig. 6c). The northern Australia hotspot matches the location and timing of peak OMI HCHO (Extended Data Fig. 6). GEOS-Chem does not capture the observed spatial distribution, instead predicting peak enhancements over eastern Australia in January and weaker enhancements to the north and south (Fig. 6c and Supplementary Fig. 7). As over the southeast United States, spatially varying NO_x biases are apparent and play a role in the above isoprene discrepancies.

Over southeastern Australia, the CrIS isoprene columns peak in January, with an ~25% decrease from January to April and a July minimum. GEOS-Chem predicts a much larger (~90%) January–April drop, with mean columns 40–95% lower than observed. In situ measurements from the Sydney Particle Study[49] support the weaker seasonality seen by CrIS (Fig. 6d). The CrIS-based source optimization shows that this modest seasonality also manifests in the underlying isoprene emissions (Supplementary Fig. 15).

Future outlook

We presented a global picture of isoprene from space, derived from CrIS radiances using an ANN. The reliability of the CrIS measurements is supported by comparisons to aircraft data and to (previously validated) optimal estimation measurements. However, more extensive validation data are needed to better quantify uncertainties and refine the measurement approach presented here.

Combining the CrIS measurements with contemporaneous HCHO observations provides a new space-based constraint on isoprene lifetimes, OH levels and emissions. The satellite-derived isoprene:HCHO column ratios support current understanding of isoprene–OH chemistry as represented in GEOS-Chem. In particular, the satellite data provide no indication of substantial missing OH recycling under high-isoprene, low-NO_x conditions. A comparison between measured and predicted isoprene columns over key hotspot regions elucidates spatial and temporal biases in modelled isoprene emissions and NO_x, which highlight the need for better mechanistic understanding of the drivers of tropical isoprene and NO_x sources.

Finally, this work lays a foundation for multi-year studies examining seasonal-to-interannual isoprene changes and their impacts on atmospheric chemistry. Supplementary Note 7 illustrates this potential by applying the CrIS ANN retrieval from 2012–2018 over Amazonia and southern Africa (Supplementary Fig. 18). Results show that the strong seasonal patterns discussed earlier persist from year to year, but also reveal interannual differences tied to temperature shifts and climate features such as El Niño. Future analyses of the full global CrIS isoprene record can therefore elucidate key drivers of interannual ecosystem variability, including drought and other disturbance, and the couplings between climate, ecosystems and atmospheric chemistry.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2664-3.

1. Guenther, A. B. et al. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model Dev.* **5**, 1471–1492 (2012).
2. Saunois, M. et al. The global methane budget 2000–2012. *Earth Syst. Sci. Data* **8**, 697–751 (2016).
3. Huang, G. L. et al. Speciation of anthropogenic emissions of non-methane volatile organic compounds: a global gridded data set for 1970–2012. *Atmos. Chem. Phys.* **17**, 7683–7701 (2017).
4. Trainer, M. et al. Models and observations of the impact of natural hydrocarbons on rural ozone. *Nature* **329**, 705–707 (1987).
5. Hewitt, C. N. et al. Ground-level ozone influenced by circadian control of isoprene emissions. *Nat. Geosci.* **4**, 671–674 (2011).
6. Mao, J. Q. et al. Ozone and organic nitrates over the eastern United States: sensitivity to isoprene chemistry. *J. Geophys. Res. Atmos.* **118**, 11256–11268 (2013).
7. Lin, Y. H. et al. Epoxide as a precursor to secondary organic aerosol formation from isoprene photooxidation in the presence of nitrogen oxides. *Proc. Natl Acad. Sci. USA* **110**, 6718–6723 (2013).
8. Bates, K. H. & Jacob, D. J. A new model mechanism for atmospheric oxidation of isoprene: global effects on oxidants, nitrogen oxides, organic products, and secondary organic aerosol. *Atmos. Chem. Phys.* **19**, 9613–9640 (2019).
9. Arneth, A. et al. Global terrestrial isoprene emission models: sensitivity to variability in climate and vegetation. *Atmos. Chem. Phys.* **11**, 8037–8052 (2011).
10. Lelieveld, J. et al. Atmospheric oxidation capacity sustained by a tropical forest. *Nature* **452**, 737–740 (2008).
11. Fuchs, H. et al. Experimental evidence for efficient hydroxyl radical regeneration in isoprene oxidation. *Nat. Geosci.* **6**, 1023–1026 (2013).
12. Feiner, P. A. et al. Testing atmospheric oxidation in an Alabama forest. *J. Atmos. Sci.* **73**, 4699–4710 (2016).
13. Rohrer, F. et al. Maximum efficiency in the hydroxyl-radical-based self-cleansing of the troposphere. *Nat. Geosci.* **7**, 559–563 (2014).
14. Bauwens, M. et al. Nine years of global hydrocarbon emissions based on source inversion of OMI formaldehyde observations. *Atmos. Chem. Phys.* **16**, 10133–10158 (2016).
15. Valin, L. C., Fiore, A. M., Chance, K. & Abad, G. G. The role of OH production in interpreting the variability of CH_2O columns in the southeast US. *J. Geophys. Res. Atmos.* **121**, 478–493 (2016).
16. Barkley, M. P. et al. Net ecosystem fluxes of isoprene over tropical South America inferred from Global Ozone Monitoring Experiment (GOME) observations of HCHO columns. *J. Geophys. Res. Atmos.* **113**, D20304 (2008).
17. Zhu, L. et al. Anthropogenic emissions of highly reactive volatile organic compounds in eastern Texas inferred from oversampling of satellite (OMI) measurements of HCHO columns. *Environ. Res. Lett.* **9**, 114004 (2014).
18. Boeke, N. L. et al. Formaldehyde columns from the Ozone Monitoring Instrument: urban versus background levels and evaluation using aircraft data and a global model. *J. Geophys. Res. Atmos.* **116**, D05303 (2011).
19. Fu, D. et al. Direct retrieval of isoprene from satellite-based infrared measurements. *Nat. Commun.* **10**, 3811 (2019).
20. Brauer, C. S. et al. Quantitative infrared absorption cross sections of isoprene for atmospheric measurements. *Atmos. Meas. Technol.* **7**, 3839–3847 (2014).
21. Razavi, A. et al. Global distributions of methanol and formic acid retrieved for the first time from the IASI/MetOp thermal infrared sounder. *Atmos. Chem. Phys.* **11**, 857–872 (2011).
22. Clarisse, L., Clerbaux, C., Dentener, F., Hurtmans, D. & Coheur, P. F. Global ammonia distribution derived from infrared satellite observations. *Nat. Geosci.* **2**, 479–483 (2009).
23. Franco, B. et al. A general framework for global retrievals of trace gases from IASI: application to methanol, formic acid, and PAN. *J. Geophys. Res. Atmos.* **123**, 13963–13984 (2018).
24. Whitburn, S. et al. A flexible and robust neural network IASI-NH_3 retrieval algorithm. *J. Geophys. Res. Atmos.* **121**, 6581–6599 (2016).
25. Warneke, C. et al. Instrumentation and measurement strategy for the NOAA SENEX aircraft campaign as part of the Southeast Atmosphere Study 2013. *Atmos. Meas. Technol.* **9**, 3063–3093 (2016).
26. Toon, O. B. et al. Planning, implementation, and scientific goals of the Studies of Emissions and Atmospheric Composition, Clouds and Climate Coupling by Regional Surveys (SEAC^{4}RS) field mission. *J. Geophys. Res. Atmos.* **121**, 4967–5009 (2016).
27. Xie, Y. et al. Understanding the impact of recent advances in isoprene photooxidation on simulations of regional air quality. *Atmos. Chem. Phys.* **13**, 8439–8455 (2013).
28. Teng, A. P., Crounse, J. D. & Wennberg, P. O. Isoprene peroxy radical dynamics. *J. Am. Chem. Soc.* **139**, 5367–5377 (2017).
29. Kim, S.-W., Barth, M. C. & Trainer, M. Impact of turbulent mixing on isoprene chemistry. *Geophys. Res. Lett.* **43**, 7701–7708 (2016).
30. De Smedt, I. et al. Algorithm theoretical baseline for formaldehyde retrievals from S5P TROPOMI and from the QA4ECV project. *Atmos. Meas. Technol.* **11**, 2395–2426 (2018).
31. Boersma, K. F. et al. Improving algorithms and uncertainty estimates for satellite NO_2 retrievals: results from the quality assurance for the essential climate variables (QA4ECV) project. *Atmos. Meas. Tech.* **11**, 6651–6678 (2018).
32. de Gouw, J. A. et al. Hydrocarbon removal in power plant plumes shows nitrogen oxide dependence of hydroxyl radicals. *Geophys. Res. Lett.* **46**, 7752–7760 (2019).
33. Wei, D. D. et al. Environmental and biological controls on seasonal patterns of isoprene above a rain forest in central Amazonia. *Agric. For. Meteorol.* **256–257**, 391–406 (2018).
34. Barkley, M. P. et al. Regulated large-scale annual shutdown of Amazonian isoprene emissions? *Geophys. Res. Lett.* **36**, L04803 (2009).
35. Alves, E. G. et al. Leaf phenology as one important driver of seasonal changes in isoprene emissions in central Amazonia. *Biogeosciences* **15**, 4019–4032 (2018).
36. Silvern, R. F. et al. Using satellite observations of tropospheric NO_2 columns to infer long-term trends in US NO_x emissions: the importance of accounting for the free tropospheric NO_2 background. *Atmos. Chem. Phys.* **19**, 8863–8878 (2019).
37. Belmonte Rivas, M. et al. OMI tropospheric NO_2 profiles from cloud-slicing: constraints on surface emissions, convective transport and lightning NO_x. *Atmos. Chem. Phys.* **15**, 13519–13553 (2015).

38. Martin, S. T. et al. Introduction: observations and modeling of the Green Ocean Amazon (GoAmazon2014/5). *Atmos. Chem. Phys.* **16**, 4785–4797 (2016).
39. Liu, Y. et al. Isoprene photochemistry over the Amazon rainforest. *Proc. Natl Acad. Sci. USA* **113**, 6125–6130 (2016).
40. Guenther, A. et al. Isoprene emission estimates and uncertainties for the Central African EXPRESSO study domain. *J. Geophys. Res. Atmos.* **104**, 30625–30639 (1999).
41. Marais, E. A. et al. Isoprene emissions in Africa inferred from OMI observations of formaldehyde columns. *Atmos. Chem. Phys.* **12**, 6219–6235 (2012).
42. Otter, L. B., Guenther, A. & Greenberg, J. Seasonal and spatial variations in biogenic hydrocarbon emissions from southern African savannas and woodlands. *Atmos. Environ.* **36**, 4265–4275 (2002).
43. Otter, L. et al. Spatial and temporal variations in biogenic volatile organic compound emissions for Africa south of the equator. *J. Geophys. Res. Atmos.* **108**, 8505 (2003).
44. Stavrakou, T. et al. Global emissions of non-methane hydrocarbons deduced from SCIAMACHY formaldehyde columns through 2003–2006. *Atmos. Chem. Phys.* **9**, 3663–3679 (2009).
45. Marais, E. A. et al. Improved model of isoprene emissions in Africa using Ozone Monitoring Instrument (OMI) satellite observations of formaldehyde: implications for oxidants and particulate matter. *Atmos. Chem. Phys.* **14**, 7693–7703 (2014).
46. Wiedinmyer, C. et al. Ozarks Isoprene Experiment (OZIE): measurements and modeling of the "isoprene volcano". *J. Geophys. Res. Atmos.* **110**, D18307 (2005).
47. Kaiser, J. et al. High-resolution inversion of OMI formaldehyde columns to quantify isoprene emission on ecosystem-relevant scales: application to the southeast US. *Atmos. Chem. Phys.* **18**, 5483–5497 (2018).
48. Hansen, D. A. et al. The southeastern aerosol research and characterization study: Part 1-overview. *J. Air Waste Manag. Assoc.* **53**, 1460–1471 (2003).
49. Emmerson, K. M. et al. Current estimates of biogenic emissions from eucalypts uncertain for southeast Australia. *Atmos. Chem. Phys.* **16**, 6997–7011 (2016).

Methods

CrIS satellite sensor

CrIS is a Fourier-transform spectrometer that was launched onboard the Suomi-NPP satellite in October 2011. A second CrIS instrument was launched onboard NOAA-20 in November 2017, and a third is planned for inclusion on JPSS-2 (expected launch in 2022). CrIS flies in a sun-synchronous orbit with 13:30 LT daytime Equator overpass. The early afternoon overpass is advantageous as it coincides with peak isoprene emissions[50] as well as with enhanced surface-atmosphere thermal contrast and vertical mixing—both of which increase the sensitivity of thermal IR sounders to near-surface absorbers. CrIS has an angular field of regard consisting of a 3 × 3 pixel array (each with a 14-km-diameter nadir footprint) and a cross-track scan width of 2,200 km, resulting in near-global coverage twice daily. The CrIS measurements have 0.625 cm^{-1} spectral resolution in the longwave IR[51], with noise characteristics (~0.04 K at 280 K) that improve substantially over other atmospheric sounders[52]. The high spectral resolution and low noise provide additional key advantages for measuring atmospheric isoprene.

GEOS-Chem simulation

We use the GEOS-Chem 3D CTM as an intercomparison platform for evaluating the isoprene estimates from CrIS, and to interpret the space-based observations in terms of isoprene emissions and chemistry. The model (v11-02e; www.geos-chem.org) employs GEOS-5 FP meteorological data from the NASA Global Modelling and Assimilation Office (GMAO), here regridded to 2° latitude × 2.5° longitude with 47 levels from the surface to 0.01 hPa. Simulations use a 10 min transport timestep (20 min for emissions and chemistry) and 1 year initialization. Model output for 12:00–15:00 LT is used for comparison with the ~13:30 LT CrIS and OMI observations.

GEOS-Chem includes detailed HO_x–NO_x–VOC–ozone–BrO_x chemistry coupled to aerosols[6,53]. The v11-02e isoprene oxidation scheme[54–56] (which is consistent with the standard v11-02c mechanism detailed by Bates and Jacob[8]) has been extensively updated to reflect recent laboratory and field-based findings, in particular for the reaction of isoprene peroxy radicals ($ISOPO_2$) with HO_2[57] and isoprene epoxides with OH[58], $ISOPO_2$ self-reaction[27], aerosol uptake of isoprene oxidation products[55] and isoprene nitrate chemistry[54,59]. $ISOPO_2$ isomerization[60–62] is treated explicitly, with oxidation and photolysis of the resulting hydroperoxyaldehydes following the current state of science[62–65] as described by Fisher et al.[54].

Along with base-case simulations using the standard (v11-02e) mechanism above, we perform sensitivity analyses using the Mini-CIM version of the reduced Caltech Isoprene Mechanism (RCIM[8,66]), implemented in GEOS-Chem v11-02c. Mini-CIM is streamlined from the parent RCIM mechanism outlined by Wennberg et al.[66] by lumping very-low-yield (<0.1% globally) isoprene oxidation products to arrive at a number of organic species and reactions comparable to what is used in current global models. Bates and Jacob[8] found global model results using Mini-CIM to be highly consistent with those using the more explicit parent mechanism (for example, the global methane lifetime difference is <0.1%), and thus recommend its use except in specialized applications involving highly functionalized, low-yield isoprene oxidation products.

An important feature of Mini-CIM is its dynamic treatment of the allylic and peroxy radicals resulting from the initial OH + isoprene addition[28,67] versus the fixed distributions used in prior mechanisms (including GEOS-Chem v11-02e). Mini-CIM also includes more intermolecular H shifts than older mechanisms, including rapid peroxy–hydroperoxy shifts[68,69] that increase low-NO OH recycling compared with GEOS-Chem v11-02e. An additional difference from our base-case simulations lies in the fact that Mini-CIM predicts more HCHO production at low-NO_x, with differences reaching approximately 20% for NO between 1 and 20 ppt (ref. [8]).

Biogenic emissions of isoprene and other VOCs are simulated using MEGANv2.1 (ref. [1]), implemented in GEOS-Chem as described by Hu et al.[70]. Global anthropogenic emissions are based on the RETRO inventory for VOCs and on EDGARv4.2[71] for NO_x, SO_x and CO; each is overwritten by regional inventories over the United States[72], Canada, Mexico[73], Europe[74] and Asia[75]. GFED4[76] is used to compute biomass burning emissions; lightning and soil NO_x emissions are from Murray et al.[77] and Hudman et al.[78], respectively.

Isoprene signal and brightness temperature difference

Isoprene has two IR absorption features (ν_{27} and ν_{28}) in the vicinity of 900 cm^{-1} that are associated with the wagging vibrational mode for each of the molecule's $=CH_2$ groups[20]. Extended Data Fig. 1a illustrates the radiance signal arising from those absorption features, plotted as the simulated difference in brightness temperature between an atmosphere with and without isoprene, assuming an isoprene profile with 5 ppb in the boundary layer and the US Standard Atmosphere[79] for interfering species. Fu et al.[19] demonstrated previously that the ν_{27} and ν_{28} features shown in Extended Data Fig. 1a are detectable from individual CrIS spectra over high-isoprene regions.

We start here from single-footprint Level 1B CrIS radiances that have been subsetted (1 of each 3 × 3 pixel array; FOV 6), cloud screened and gridded to 0.5° latitude × 0.625° longitude. The ΔT_b values are then calculated as the difference between off-peak (mean of the spectral points at 894.375 and 895 cm^{-1}) and on-peak (mean of the spectral points at 893.125 and 893.75 cm^{-1}) T_b values at the ν_{28} feature.

Cloud screening is based on the observed difference between the 900 cm^{-1} brightness temperature and the surface skin temperature. We simulate this difference for clear-sky conditions as a function of water vapour column density (solid black line in Extended Data Fig. 7a) using the Line-by-Line Radiative Transfer Model[80,81] and employ a conservative linear approximation (solid red line in Extended Data Fig. 7a) to screen the observations. Temperature and water vapour information is from MERRA-2 reanalysis[82] and interpolated to the time of CrIS overpass. We find good spatial correspondence between the location of our cloud-screened pixels and cloud flags derived from other space-borne sensors such as VIIRS and MODIS.

Given the demonstrated importance of careful cloud screening for optimal estimation isoprene retrievals from CrIS[19], we test the sensitivity of our results to cloud effects by employing a less stringent (by 2 K) brightness temperature threshold (dashed red line in Extended Data Fig. 7a). The results of this test are summarized in Extended Data Fig. 7b, c, and show that the resulting ΔT_b and isoprene changes are generally less than 15%, and less than 5% for enhanced isoprene levels. This suggests that the uncertainty in results presented here is not dominated by cloud effects.

Extended Data Figure 1a shows that other atmospheric species (specifically water vapour, nitric acid, ammonia and CFC-12) also have absorption features in the vicinity of the ν_{27} and ν_{28} isoprene peaks. We specifically chose to use ν_{28} in computing ΔT_b as it is the stronger of the two bands and less subject to such interferences. Nevertheless, variability in these other atmospheric species (and in factors such as surface–atmosphere thermal contrast, surface elevation and satellite viewing angle) can still affect the ΔT_b–isoprene relationship[19], and are therefore accounted for in the estimation process described in the following section.

While other biogenically derived VOCs with terminal $=CH_2$ groups may also absorb in the vicinity of the isoprene peaks, Fu et al.[19] showed that the relevant primary biogenic species (including monoterpenes) with published absorption cross-sections have much weaker absorption signals (<0.01 K) than isoprene at ν_{28}. Since we focus here on isoprene hotspots, we assume such effects to be minor for our analysis. Relevant absorption cross-sections for key non-HCHO isoprene oxidation products (methyl vinyl ketone, methacrolein, isoprene hydroxyhydroperoxides) have not been reported, but available analogues indicate that their

spectral impact is likewise minor for analyses here (Supplementary Fig. 19). See Supplementary Note 8 for further discussion.

Extended Data Table 1 shows spatial correlations between the resulting CrIS ΔT_b measurements and simulated isoprene columns from the GEOS-Chem CTM over key source regions. Here and below, all satellite–model comparisons reflect monthly mean values at the ~13:30 LT CrIS overpass with daily cloud screening. Correlations span $r = 0.43–0.72$. For comparison, Hu et al.[70] report $r = 0.5–0.7$ between simulated and measured isoprene in the midwest United States. A model–aircraft comparison over the southeast United States yields similar correlations (below). The CrIS ΔT_b values thus spatially correlate with isoprene predictions over known source regions to a degree commonly found for model–measurement comparisons of isoprene itself.

ANN training and forward prediction

We describe here a supervised feed-forward (that is, non-cyclic) ANN[83] to derive isoprene columns from the CrIS ΔT_b observations. The approach employs a multilayer perceptron with training via Levenberg–Marquardt backpropagation[84] to account for the interfering effects mentioned above based on contemporaneous observations of other relevant surface and atmospheric properties.

Given a set of input variables $\mathbf{x}$ (in our case, ΔT_b and related parameters summarized in Extended Data Table 2), an ANN can be used to approximate an output $f(\mathbf{x})$ (in our case, $\Omega_{isoprene}$) that depends on $\mathbf{x}$ in an unknown and possibly nonlinear way. This approximation occurs via a transfer function, $Y(W, \mathbf{x})$, where W represents the weights of the function Y.

The weights are determined here with a synthetic dataset, constructed based on a full year of simulated radiances from the Earth Limb and Nadir Operational Retrieval (ELANOR) model[85], which also serves as the operational forward model for the Tropospheric Emission Spectrometer (TES). ELANOR model inputs include temperature and water vapour profiles (using assimilated meteorological data from NASA GMAO) and climatological non-isoprene trace gas profiles (from the MOZART CTM[86]). Isoprene profiles are taken from daily mid-afternoon (12:00–15:00 LT) GEOS-Chem predictions with 100% (1σ) Gaussian noise applied. We then apply global sampling (afternoon overpass, following the along-track separation of measurements from the global sampling strategy of TES[87], land locations only) to arrive at a representative input dataset of appropriate size for ANN training. Finally, the resulting radiances are simulated (using temperature-dependent isoprene absorption look-up tables) for three satellite viewing angles (selected randomly for each location). The full synthetic dataset comprises ~165,000 simulated spectra, from which we compute ΔT_b as above.

We then train the ANN to predict isoprene column densities based on six predictors (each taken as a firm constraint): ΔT_b, water vapour column density (Ω_{H2O}), column nitric acid density (Ω_{HNO3}), thermal contrast (taken as the difference between the surface skin and 2 m air temperatures), surface pressure and satellite viewing angle. Alternate ANNs accounting for other potential interferents (such as CFCs and ammonia) were tested but ultimately discarded as they contributed little additional power to the isoprene predictions. No location-specific information is included in the training: the network thus describes the general global relationship between ΔT_b, isoprene columns and associated factors that is mechanistically defined by the underlying spectroscopy. This is a key distinction from optimal estimation retrievals, which incorporate varying amounts of prior information depending on the location-specific sensitivity.

We assessed multiple network architectures and found the best performance for a three-layer model containing two (six- and three-neuron) hidden layers and one (single-neuron) output layer using hyperbolic tangent (sigmoid) and linear transfer functions, respectively. The training occurs on ten random extractions of the synthetic dataset (after clustering to ensure representative sampling across the full range of isoprene column densities), with each extraction subsetted for training (50%), validation (30%) and testing (20%). The validation subset is used to determine when training can cease, and the testing subset is used subsequently to independently confirm network performance. Output from the resulting ten networks are then averaged to provide the final ANN prediction.

Finally, we apply the trained ANN to the space-borne CrIS ΔT_b measurements to derive global isoprene distributions for January, April, July and October 2013. Temperature and water vapour data are taken from the MERRA-2 reanalysis[82] and interpolated to the CrIS overpass time, whereas nitric acid column observations are from the CrIS CLIMCAPS[88] product. All input variables are cloud-screened as described above before calculation of the gridded (2° × 2.5°) 13:30 LT monthly mean. Fewer than 1% of the employed input variables fall outside the range used for ANN training (none of which occur over isoprene source regions), confirming that our training set is well-generalized.

Unlike a conventional optimal estimation retrieval, the ANN-based approach does not provide an estimate of the measurement vertical sensitivity (that is, averaging kernel) and associated uncertainty for every individual location. However, the ANN training statistics provide a quantification of the overall network performance, and therefore of the expected uncertainties for isoprene column abundances inferred from CrIS data. We find here that the six-predictor ANN can reproduce 93% of the variance in the isoprene total columns across the full synthetic dataset (Extended Data Fig. 7d). The performance of each of the ten networks relative to the independent testing set is similar ($r^2 = 0.92–0.93$, slopes ~1.0). This explanatory skill is lost when ΔT_b is withheld from training ($r^2 = 0.28$; Extended Data Fig. 7e)—confirming that the predictive power of the ANN is driven by the isoprene spectral signal rather than by the ancillary variables.

The relative uncertainty of the ANN predictions varies as a function of both isoprene amount and thermal contrast (Extended Data Fig. 7f). For enhanced isoprene columns (>1×10^{16} molecules per cm^2) the prediction uncertainty is typically less than 30%, even with very low thermal contrast. Uncertainty increases for lower isoprene amounts, exceeding 50% for columns less than 2×10^{15} molecules per cm^2, and for columns less than 5×10^{15} molecules per cm^2 at low thermal contrast (0–5 K; Extended Data Fig. 1c shows thermal contrast maps for January, April, July and October). These can be considered limits of detection for the 13:30 LT monthly mean isoprene columns derived from CrIS.

The statistical performance of the ANN as summarized above does not necessarily represent the full uncertainty of the CrIS isoprene measurements, as other factors (for example, cross-section or radiative transfer errors, uncertainties in ancillary datasets used for water vapour, temperature and HNO_3, uncertainties in the vertical profiles of isoprene used to train the ANN, residual cloud impacts) may also contribute. We therefore evaluate the CrIS isoprene columns using (1) the previously published and validated optimal estimation retrievals and (2) independent atmospheric measurements, as described below and in the main text.

CrIS evaluation via aircraft–model intercomparison

Direct evaluation of the CrIS isoprene measurements is difficult due to lack of either ground-based isoprene column observations in isoprene hotspot regions, or a statistically sufficient ensemble of full airborne profiles over isoprene source regions at the satellite overpass time. Instead, we perform an indirect validation (Fig. 2b–d) using measurements from two aircraft campaigns over the southeast United States: SENEX (Southeast Nexus; 27 May–10 July 2013; ref. [25]) and SEAC^{4}RS (Studies of Emissions and Atmospheric Composition, Clouds and Climate Coupling by Regional Surveys; 1 August–23 September 2013; ref. [26]). In each case, we employ the GEOS-Chem model as an intercomparison platform to quantify the level of consistency between CrIS and the in situ aircraft data. Since any model isoprene bias should manifest in a consistent way relative to independent observational datasets for

the same region and time period, the consistency between the CrIS/GEOS-Chem regression and the aircraft/GEOS-Chem regression reflects the agreement between the CrIS and in situ isoprene datasets[89,90].

To perform this intercomparison, we sample the model at the time and location of the aircraft measurements (which are restricted to ±2 h from the CrIS overpass time). Results discussed in the main text are aggregated to the model resolution and averaged vertically for each campaign by calculating a density-weighted mean boundary layer (pressure $P > 800$ hPa) number density for each latitude × longitude grid cell.

OMI HCHO and NO_2 data

We use here the Quality Assurance for Essential Climate Variables (QA4ECV) version 1.0 Level 2 HCHO product from the OMI satellite sensor[29,91]. OMI is a near-ultraviolet–visible spectrometer onboard NASA's EOS Aura satellite, which has an Equator overpass time (13:40 LT) close to that of Suomi-NPP. The HCHO slant column density is determined via fitting of OMI radiances and subsequently converted to vertical column densities using a modelled shape factor. The QA4ECV retrieval uses a single, extended fitting interval (328.5–359.0 nm), whereas the precursor BIRA HCHO retrieval employed a smaller window with prefits for O_2–O_2 and BrO slant columns. Although the QA4ECV data have yet to be fully validated, recent work has demonstrated their improved performance over the earlier BIRA retrieval[92]. Zhu et al.[93] previously found the BIRA v14 HCHO retrieval to exhibit a 12% low bias (with use of an accurate shape factor) relative to aircraft measurements, and subsequent analysis has supported these findings[94]. We find here that a global QA4ECV versus BIRA v14 comparison for the timeframe of our analysis yields a slope of 1.1–1.4 (0.9–1.8 over our targeted subregions), and we therefore do not apply any bias correction to the QA4ECV HCHO data. Repeating our analysis using instead the bias-corrected BIRA v14 dataset (Supplementary Figs. 20–22) leads to no substantive differences in our core results.

Standard data processing and screening procedures are followed. We restrict the data to solar zenith angle <70° and cloud fraction <0.4. The OMI data are then gridded to the 2° × 2.5° GEOS-Chem resolution. For all comparisons the model is sampled according to the OMI HCHO observation operator at the time and location of the satellite overpass.

Tropospheric NO_2 column data are from the OMI QA4ECV v1.1 monthly NO_2 product[31,95]. The QA4ECV retrieval employs updated NO_2 spectral fitting that accounts for liquid water absorption and includes an intensity offset correction[31]. This improves the quality of the product, particularly over clear-sky ocean locations[92]. OMI QA4ECV tropospheric NO_2 columns exhibited good agreement (bias = −2% and root-mean-square difference = 16%) with ground-based column measurements in China[31]. Comparisons in this work are performed with respect to monthly mean GEOS-Chem tropospheric NO_2 columns sampled at the time of the satellite overpass, with no observation operator applied.

Data availability

The CrIS Level 1B data used in this work are publicly available at https://snpp-sounder.gesdisc.eosdis.nasa.gov/data/SNPP_Sounder_Level1/SNPPCrISL1BNSR.1/. The isoprene column data employed in this work are available at https://doi.org/10.13020/v959-dr15. The airborne data are publicly available for SENEX at http://esrl.noaa.gov/csd/projects/senex/ and for SEAC^{4}RS at http://www-air.larc.nasa.gov/missions/seac4rs/index.html. OMI QA4ECV HCHO and NO_2 data are publicly available at http://www.qa4ecv.eu/ecvs.

Code availability

GEOS-Chem model code is publicly available at http://www.geos-chem.org. The LBLRTM[80,81], which is used to calculate the molecular absorption look-up tables employed in ELANOR[85], is publicly available at http://rtweb.aer.com/lblrtm.html.

50. Guenther, A. B. & Hills, A. J. Eddy covariance measurement of isoprene fluxes. *J. Geophys. Res. Atmos.* **103**, 13145–13152 (1998).
51. Han, Y. et al. Suomi NPP CrIS measuremets, sensor data record algorithm, calibration and validation activities, and record data quality. *J. Geophys. Res. Atmos.* **118**, 12734–12748 (2013).
52. Zavyalov, V. et al. Noise performance of the CrIS instrument. *J. Geophys. Res. Atmos.* **118**, 13108–13120 (2013).
53. Millet, D. B. et al. A large and ubiquitous source of atmospheric formic acid. *Atmos. Chem. Phys.* **15**, 6283–6304 (2015).
54. Fisher, J. A. et al. Organic nitrate chemistry and its implications for nitrogen budgets in an isoprene- and monoterpene-rich atmosphere: constraints from aircraft (SEAC^{4}RS) and ground-based (SOAS) observations in the Southeast US. *Atmos. Chem. Phys.* **16**, 5969–5991 (2016).
55. Marais, E. A. et al. Aqueous-phase mechanism for secondary organic aerosol formation from isoprene: application to the southeast United States and co-benefit of SO_2 emission controls. *Atmos. Chem. Phys.* **16**, 1603–1618 (2016).
56. Travis, K. R. et al. Why do models overestimate surface ozone in the Southeast United States? *Atmos. Chem. Phys.* **16**, 13561–13577 (2016).
57. Liu, Y. J., Herdlinger-Blatt, I., McKinney, K. A. & Martin, S. T. Production of methyl vinyl ketone and methacrolein via the hydroperoxyl pathway of isoprene oxidation. *Atmos. Chem. Phys.* **13**, 5715–5730 (2013).
58. Bates, K. H. et al. Gas phase production and loss of isoprene epoxydiols. *J. Phys. Chem. A* **118**, 1237–1246 (2014).
59. Jacobs, M. I., Burke, W. J. & Elrod, M. J. Kinetics of the reactions of isoprene-derived hydroxynitrates: gas phase epoxide formation and solution phase hydrolysis. *Atmos. Chem. Phys.* **14**, 8933–8946 (2014).
60. Crounse, J. D., Paulot, F., Kjaergaard, H. G. & Wennberg, P. O. Peroxy radical isomerization in the oxidation of isoprene. *Phys. Chem. Chem. Phys.* **13**, 13607–13613 (2011).
61. Peeters, J., Nguyen, T. L. & Vereecken, L. HO_x radical regeneration in the oxidation of isoprene. *Phys. Chem. Chem. Phys.* **11**, 5935–5939 (2009).
62. Wolfe, G. M. et al. Photolysis, OH reactivity and ozone reactivity of a proxy for isoprene-derived hydroperoxyenals (HPALDs). *Phys. Chem. Chem. Phys.* **14**, 7276–7286 (2012).
63. Peeters, J. & Muller, J. F. HO_x radical regeneration in isoprene oxidation via peroxy radical isomerisations. II: experimental evidence and global impact. *Phys. Chem. Chem. Phys.* **12**, 14227–14235 (2010).
64. Stavrakou, T., Peeters, J. & Muller, J. F. Improved global modelling of HO_x recycling in isoprene oxidation: evaluation against the GABRIEL and INTEX-A aircraft campaign measurements. *Atmos. Chem. Phys.* **10**, 9863–9878 (2010).
65. Squire, O. J. et al. Influence of isoprene chemical mechanism on modelled changes in tropospheric ozone due to climate and land use over the 21st century. *Atmos. Chem. Phys.* **15**, 5123–5143 (2015).
66. Wennberg, P. O. et al. Gas-phase oxidation of isoprene and its major oxidation products. *Chem. Rev.* **118**, 3337–3390 (2018).
67. Peeters, J. et al. Hydroxyl radical recycling in isoprene oxidation driven by hydrogen bonding and hydrogen tunneling: the upgraded LIM1 mechanism. *J. Phys. Chem. A* **118**, 8625–8643 (2014).
68. Jørgensen, S. et al. Rapid hydrogen shift scrambling in hydroperoxy-substituted organic peroxy radicals. *J. Phys. Chem. A* **120**, 266–275 (2016).
69. Møller, K. H. et al. The importance of peroxy radical hydrogen-shift reactions in atmospheric isoprene oxidation. *J. Phys. Chem. A* **123**, 920–932 (2019).
70. Hu, L. et al. Isoprene emissions and impacts over an ecological transition region in the US Upper Midwest inferred from tall tower measurements. *J. Geophys. Res. Atmos.* **120**, 3553–3571 (2015).
71. *Emission Database for Global Atmospheric Research (EDGAR), Release Version 4.2* (European Commission (EC) Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL), 2011); http://edgar.jrc.ec.europa.eu.
72. *2011 National Emissions Inventory (NEI) Data* (EPA, 2015); http://www.epa.gov/air-emissions-inventories/2011-national-emissions-inventory-nei-data.
73. Kuhns, H., Green, M. & Etyemezian, V. *Big Bend Regional Aerosol and Visibility Observational (BRAVO) Study Emissions Inventory* (DRI, 2003).
74. Auvray, M. & Bey, I. Long-range transport to Europe: seasonal variations and implications for the European ozone budget. *J. Geophys. Res. Atmos.* **110**, D11303 (2005).
75. Li, M. et al. MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. *Atmos. Chem. Phys.* **17**, 935–963 (2017).
76. van der Werf, G. R. et al. Global fire emissions estimates during 1997-2016. *Earth Syst. Sci. Data* **9**, 697–720 (2017).
77. Murray, L. T. et al. Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data. *J. Geophys. Res. Atmos.* **117**, D20307 (2012).
78. Hudman, R. C. et al. Steps towards a mechanistic model of global nitric oxide emissions: implementation and space-based constraints. *Atmos. Chem. Phys.* **12**, 7779–7795 (2012).
79. NASA *U.S. Standard Atmosphere, 1976* Report No. NASA-TM-X-74335 (NASA, 1976).
80. Clough, S. A. et al. Atmospheric radiative transfer modeling: a summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transfer* **91**, 233–244 (2005).
81. Alvarado, M. J. et al. Performance of the Line-By-Line Radiative Transfer Model (LBLRTM) for temperature, water vapor, and trace gas retrievals: recent updates evaluated with IASI case studies. *Atmos. Chem. Phys.* **13**, 6687–6711 (2013).
82. Gelaro, R. et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454 (2017).
83. Blum, E. K. & Li, L. K. Approximation-theory and feedforward networks. *Neural Netw.* **4**, 511–515 (1991).
84. Hagan, M. T. & Menhaj, M. B. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Netw.* **5**, 989–993 (1994).
85. Clough, S. A. et al. Forward model and Jacobians for tropospheric emission spectrometer retrievals. *IEEE Trans. Geosci. Remote Sens.* **44**, 1308–1323 (2006).

86. Emmons, L. K. et al. Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4). *Geosci. Model Dev.* **3**, 43–67 (2010).
87. Beer, R. TES on the Aura mission: scientific objectives, measurements, and analysis overview. *IEEE Trans. Geosci. Remote Sens.* **44**, 1102–1105 (2006).
88. Smith, N. & Barnet, C. D. Uncertainty characterization and propagation in the Community Long-Term Infrared Microwave Combined Atmospheric Product System (CLIMCAPS). *Remote Sens.* **11**, 1227 (2019).
89. Wells, K. C. et al. Tropospheric methanol observations from space: retrieval evaluation and constraints on the seasonality of biogenic emissions. *Atmos. Chem. Phys.* **12**, 5897–5912 (2012).
90. Chaliyakunnel, S., Millet, D. B., Wells, K. C., Cady-Pereira, K. E. & Shephard, M. W. A large underestimate of formic acid from tropical fires: constraints from space-borne measurements. *Environ. Sci. Technol.* **50**, 5631–5640 (2016).
91. De Smedt, I. et al. *QA4ECV HCHO Tropospheric Column Data from OMI (Version 1.1)* (KNMI, 2017); https://doi.org/10.18758/71021031.
92. Zara, M. et al. Improved slant column density retrieval of nitrogen dioxide and formaldehyde for OMI and GOME-2A from QA4ECV: intercomparison, uncertainty characterisation, and trends. *Atmos. Meas. Technol.* **11**, 4033–4058 (2018).
93. Zhu, L. et al. Observing atmospheric formaldehyde (HCHO) from space: validation and intercomparison of six retrievals from four satellites (OMI, GOME2A, GOME2B, OMPS) with SEAC^{4}RS aircraft observations over the southeast US. *Atmos. Chem. Phys.* **16**, 13477–13490 (2016).
94. Shen, L. et al. The 2005-2016 trends of formaldehyde columns over China observed by satellites: increasing anthropogenic emissions of volatile organic compounds and decreasing agricultural fire emissions. *Geophys. Res. Lett.* **46**, 4468–4475 (2019).
95. Boersma, K. F. et al. QA4ECV NO2 Tropospheric and Stratospheric Vertical Column Data from GOME-2A (Version 1.1) (KNMI, 2017); https://doi.org/10.21944/qa4ecv-no2-gome2a-v1.1.

Acknowledgements This work was supported by the NASA Atmospheric Composition Modeling and Analysis Program (Grant Number NNX17AF61G) and by the Minnesota Supercomputing Institute. We thank D. Fu for providing optimal estimation isoprene retrievals over Amazonia and input on this manuscript; C. Barnet, E. Manning and R. Monarrez for providing CLIMCAPS HNO_3 retrievals; M. Alvarado, K. Cady-Pereira, D. Gombos, J. Hegarty and I. Strickland for generating and testing isoprene absorption look-up tables employed here; and E. Edgerton for providing isoprene data from the SouthEastern Aerosol Research and CHaracterization (SEARCH) network. The SEARCH network was sponsored by the Southern Company and the Electric Power Research Institute. Isoprene measurements aboard the NASA DC-8 during SEAC^{4}RS were supported by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit) through the Austrian Space Applications Programme (ASAP) of the Austrian Research Promotion Agency (FFG). T. Mikoviny is acknowledged for his support during SEAC^{4}RS. We thank S. Springston for GoAmazon T3 data, which were supported by the ARM Climate Research Facility, the Central Office of the Large-Scale Biosphere Atmosphere Experiment in Amazonia (LBA), the Instituto Nacional de Pesquisas da Amazonia (INPA) and the Universidade do Estado do Amazonia (UEA). Part of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to NASA.

Author contributions D.B.M. planned the project and oversaw the scientific interpretation. K.C.W. performed the ΔT_b calculations, ANN training, CrIS isoprene retrievals and evaluation, lifetime calculations, emission optimization, and synthesis of results for major source regions. V.H.P. performed radiative transfer model simulations and provided guidance with CrIS data analysis. M.J.D. assisted with ANN training and application. D.B.M. and K.C.W. conducted the GEOS-Chem model simulations. K.H.B. worked on the development of RCIM and Mini-CIM and incorporated them into GEOS-Chem. J.A.d.G., M.G., C.W. and A.W. carried out the aircraft measurements used for evaluation. J.D.F. provided ground-based isoprene measurements from Amazonia. K.C.W. and D.B.M. wrote the manuscript. All authors reviewed and commented on the paper.

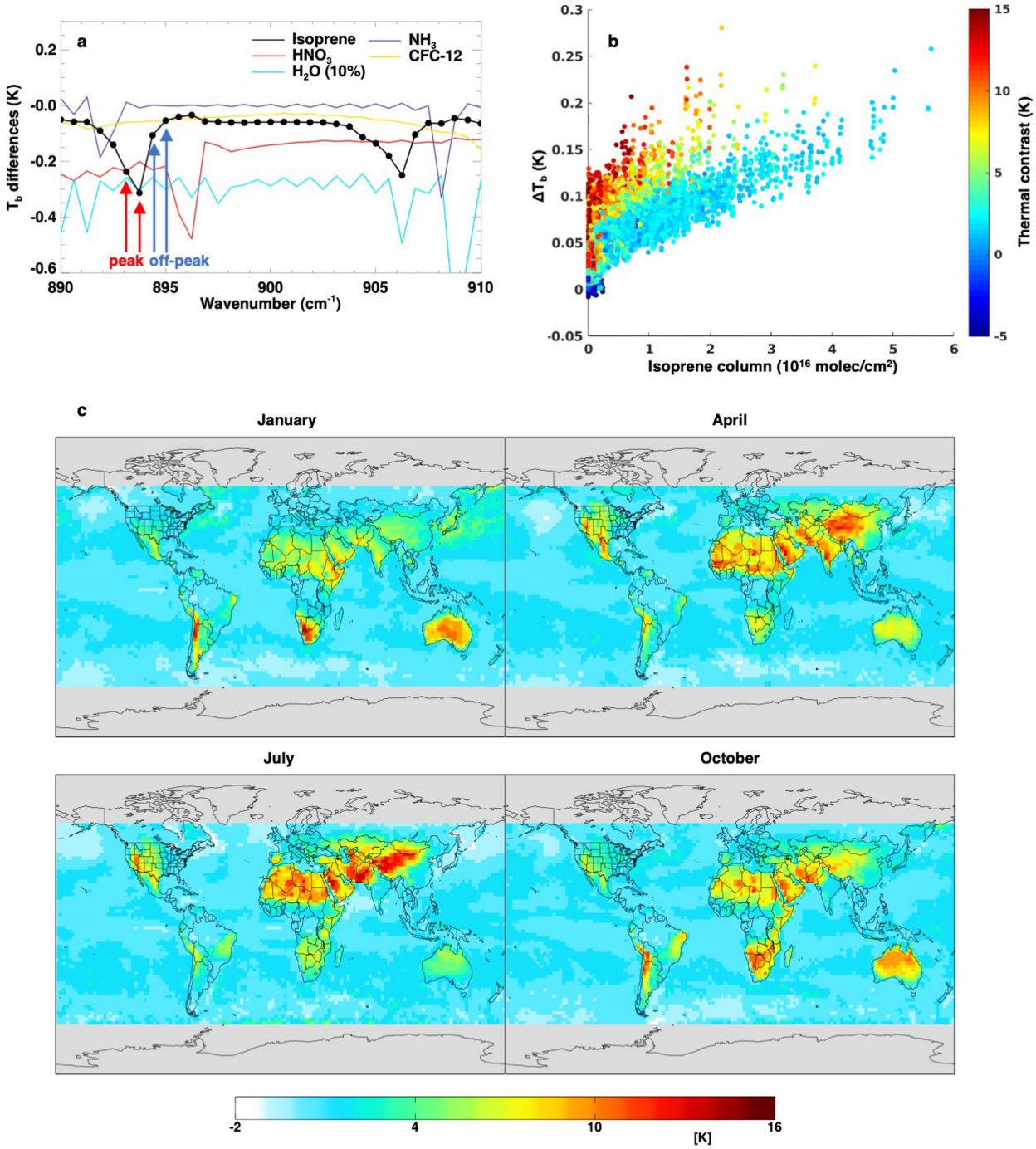
Competing interests The authors declare no competing interests.

Additional information

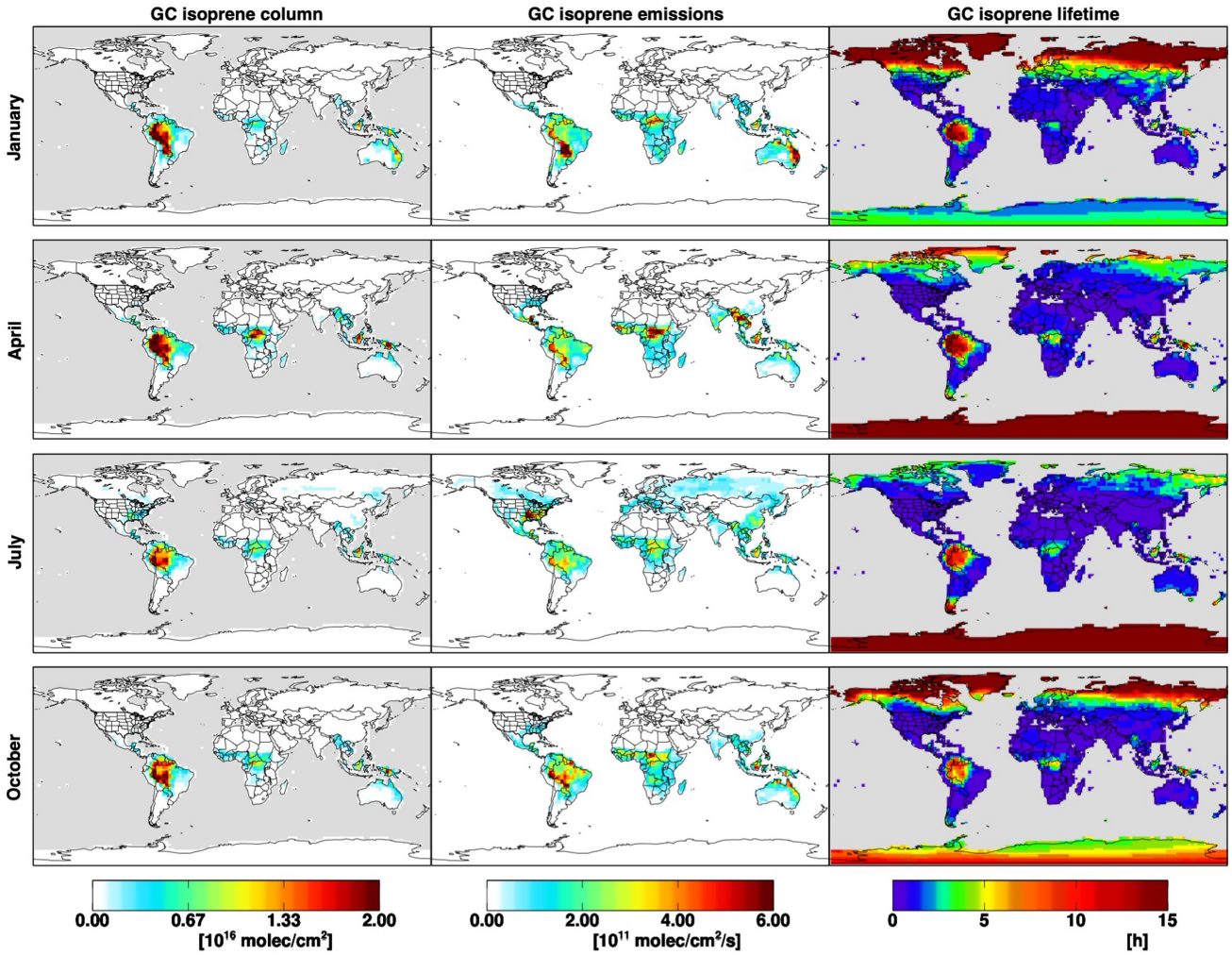
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2664-3.

Correspondence and requests for materials should be addressed to D.B.M.

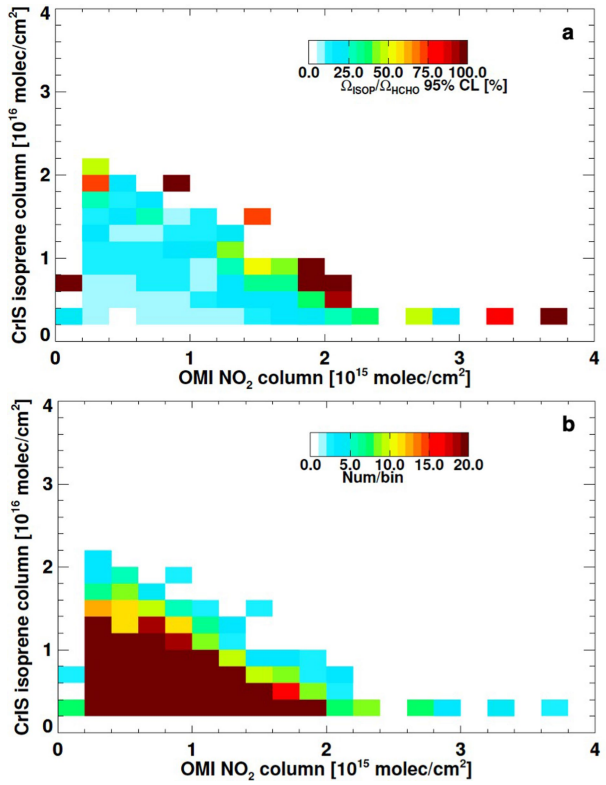
Peer review information *Nature* thanks Klaas Boersma, Jean-Francois Müller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



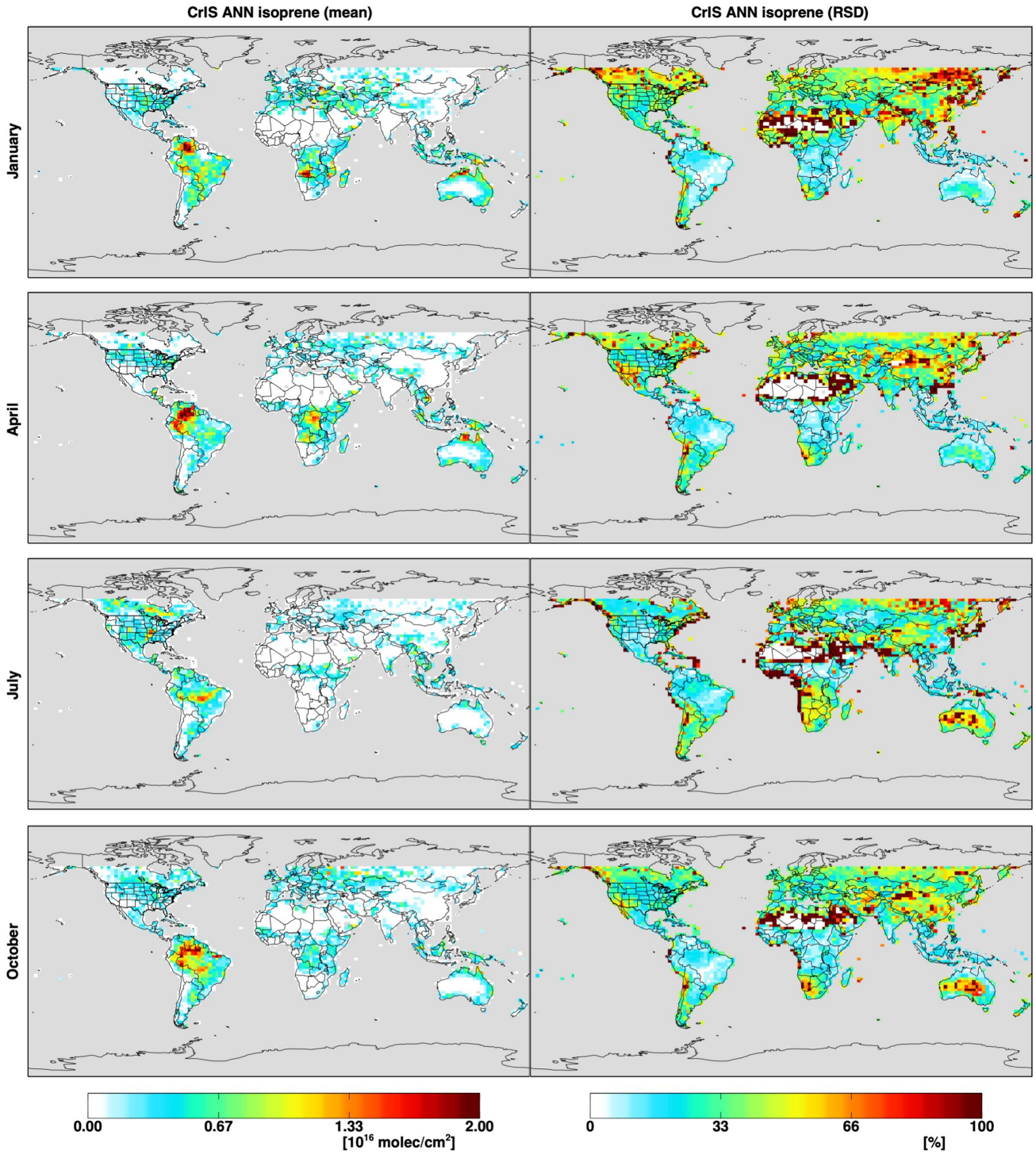
Extended Data Fig. 1 | Simulated spectral signals near 900 cm^{-1} for the CrIS sensor. a, Brightness temperature (T_b) difference for simulated spectra with and without isoprene (black), nitric acid (red), ammonia (violet) and CFC-12 (yellow) and a 10% perturbation in water vapour (cyan). Red and blue arrows indicate the ν_{28} on-peak and off-peak spectral points used to calculate ΔT_b. Simulations were performed with LBLRTM[80,81] for an isoprene profile with 5 ppb in the boundary layer ($P > 800$ hPa) that decays exponentially aloft, and US standard atmosphere profiles of temperature, water vapour and nitric acid[79]. **b**, Relationship between ΔT_b and isoprene column density, shaded by thermal contrast, for the full synthetic dataset used in this work. **c**, Global distribution of surface-atmosphere thermal contrast at the time of the CrIS overpass. Maps are derived from time-interpolated GMAO temperatures for January, April, July and October.



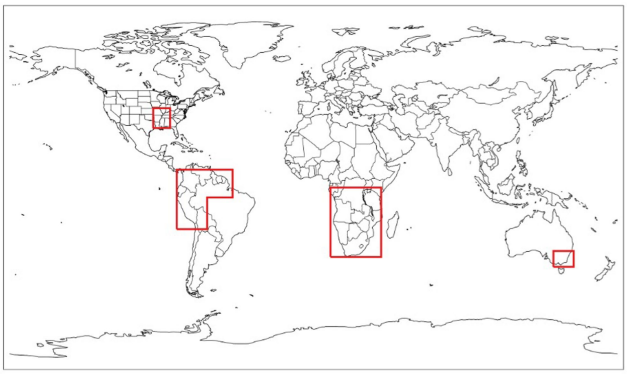
Extended Data Fig. 2 | Global distribution of isoprene columns, emissions and lifetime as predicted by GEOS-Chem. Predicted columns (left), emissions (middle) and lifetime ($z < 500$ m; right) are shown at 13:30 LT for January, April, July and October 2013 (top to bottom).



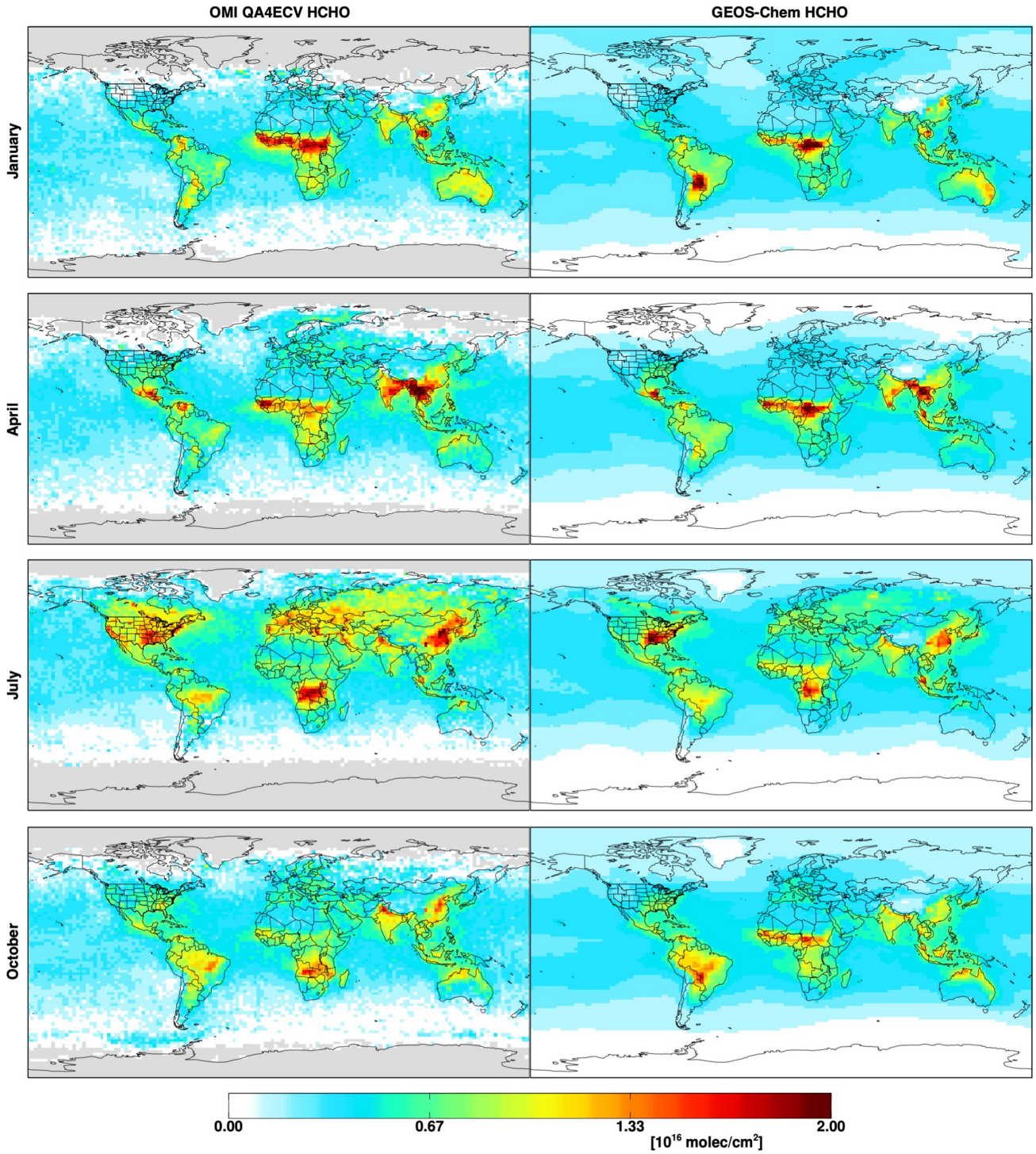
Extended Data Fig. 3 | Statistical uncertainty in the global distribution of monthly mean isoprene:HCHO ratios as a function of isoprene and NO_x regime. a, Relative 95% confidence interval in the mean ratio for each isoprene and tropospheric NO_2 bin. **b**, Number of observations in each bin.



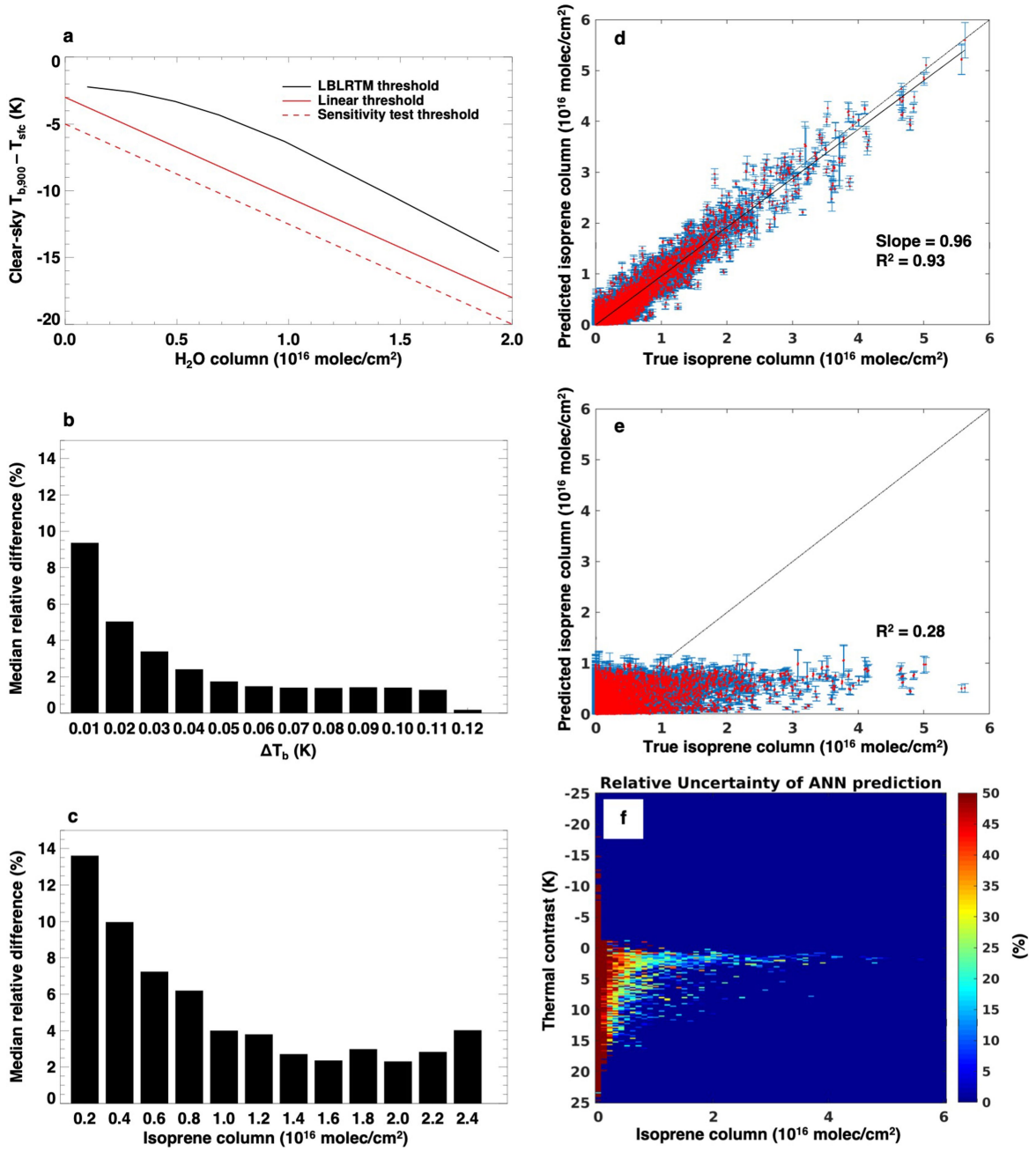
Extended Data Fig. 4 | Global distribution of isoprene column densities derived from CrIS. Plotted are the mean (left) and relative standard deviation (right) across the 10 ANNs for January, April, July and October 2013 (top to bottom).



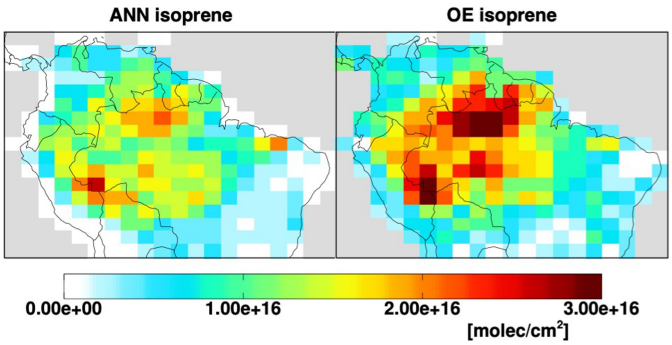
Extended Data Fig. 5 | Boundaries of the four regions examined in the seasonal bar plots shown in Figs. 5, 6.



Extended Data Fig. 6 | Measured and simulated HCHO columns. Plotted are the HCHO columns measured by OMI (left) and simulated by GEOS-Chem (right) at ~13:30 LT for January, April, July and October 2013 (top to bottom).



Extended Data Fig. 7 | CrIS cloud screening and ANN performance. **a**, Function used for cloud screening CrIS L1B data before ΔT_b calculation. The black line shows the modelled clear-sky difference between the 900 cm^{-1} brightness temperature and surface skin temperature, as a function of water vapour column density (calculated using LBLRTM[80,81]). The solid red line is the linear approximation used here, and the dashed red line represents a less stringent threshold used to test the sensitivity of the results to our cloud screening approach. **b**, **c**, Sensitivity of the CrIS brightness temperature differences (**b**) and isoprene columns (**c**) to cloud screening. Data shown represent the median relative differences between the base-case results (derived using the solid red line in **a**) and those derived using the less stringent cloud screening threshold (dashed red line in **a**). **d**, **e**, Scatterplots of the predicted versus true isoprene columns for the six-predictor ANN (**d**) and an ANN in which ΔT_b is withheld as a predictor variable (**e**). Red dots show the mean of the ten ANN predictions, and blue error bars show the standard deviation across the predictions. **f**, The relative uncertainty (based on the difference between the mean ANN predicted value and the true value) for the six-predictor ANN, binned as a function of thermal contrast and isoprene column density.



Extended Data Fig. 8 | CrIS isoprene measurements over Amazonia. The maps were derived using ANN- (left) and optimal estimation- (right) based approaches. Data are shown for September 2014 and displayed as absolute columns.

Extended Data Table 1 | Spatial correlation between monthly mean CrIS ΔT_b and monthly mean 13:30 LT isoprene columns predicted by GEOS-Chem at 2° × 2.5° resolution for select regions

Region	Month	ΔT_b:GEOS-Chem isoprene correlation, r	# data points
Australia	January	0.54	323
Central Africa	April	0.43	357
Southeast United States	July	0.72	90
Amazonia	October	0.57	340

Extended Data Table 2 | Data sources for the six input parameters used for ANN training and retrievals

Input parameter	Source for training set	Source for ANN-based retrieval
ΔT_b	ELANOR simulation	CrIS L1B radiances
H_2O vapor column	Assimilated meteorology (GMAO; TES-like sampling)	Assimilated meteorology (GMAO; CrIS collocation)
HNO_3 column	MOZART CTM	CrIS CLIMCAPS
Thermal contrast	Assimilated meteorology (GMAO; TES-like sampling)	Assimilated meteorology (GMAO; CrIS collocation)
Pressure	Assimilated meteorology (GMAO; TES-like sampling)	Assimilated meteorology (GMAO; CrIS collocation)
Satellite view angle	Randomly defined	CrIS satellite pointing angle

Article

The lithospheric-to-lower-mantle carbon cycle recorded in superdeep diamonds

https://doi.org/10.1038/s41586-020-2676-z

Received: 7 January 2020

Accepted: 14 July 2020

Published online: 9 September 2020

Check for updates

M. E. Regier[1]✉, D. G. Pearson[1], T. Stachel[1], R. W. Luth[1], R. A. Stern[1] & J. W. Harris[2]

The transport of carbon into Earth's mantle is a critical pathway in Earth's carbon cycle, affecting both the climate and the redox conditions of the surface and mantle. The largest unconstrained variables in this cycle are the depths to which carbon in sediments and altered oceanic crust can be subducted and the relative contributions of these reservoirs to the sequestration of carbon in the deep mantle[1]. Mineral inclusions in sublithospheric, or 'superdeep', diamonds (derived from depths greater than 250 kilometres) can be used to constrain these variables. Here we present oxygen isotope measurements of mineral inclusions within diamonds from Kankan, Guinea that are derived from depths extending from the lithosphere to the lower mantle (greater than 660 kilometres). These data, combined with the carbon and nitrogen isotope contents of the diamonds, indicate that carbonated igneous oceanic crust, not sediment, is the primary carbon-bearing reservoir in slabs subducted to deep-lithospheric and transition-zone depths (less than 660 kilometres). Within this depth regime, sublithospheric inclusions are distinctly enriched in ^{18}O relative to eclogitic lithospheric inclusions derived from crustal protoliths. The increased ^{18}O content of these sublithospheric inclusions results from their crystallization from melts of carbonate-rich subducted oceanic crust. In contrast, lower-mantle mineral inclusions and their host diamonds (deeper than 660 kilometres) have a narrow range of isotopic values that are typical of mantle that has experienced little or no crustal interaction. Because carbon is hosted in metals, rather than in diamond, in the reduced, volatile-poor lower mantle[2], carbon must be mobilized and concentrated to form lower-mantle diamonds. Our data support a model in which the hydration of the uppermost lower mantle by subducted oceanic lithosphere destabilizes carbon-bearing metals to form diamond, without disturbing the ambient-mantle stable-isotope signatures. This transition from carbonate slab melting in the transition zone to slab dehydration in the lower mantle supports a lower-mantle barrier for carbon subduction.

The first seismological images of subducted oceanic lithosphere penetrating the 660-km mantle discontinuity provided evidence for the circulation of some upper-mantle material into the lower mantle[3]. Nevertheless, the depths at which volatiles are lost from the slab as it subducts into the deep convecting mantle remain poorly understood. Diamonds are unique windows into this environment in that they directly sample the elemental and isotopic compositions present at these depths. As high-temperature fractionation[4] cannot account for all of the isotopic variability observed in diamonds, the ^{13}C-depleted signatures of some lithospheric-to-transition-zone diamonds are typically interpreted to reflect the deep subduction of sediments, which are rich in ^{13}C-depleted, reduced organic carbon[5,6]. This idea has garnered much attention, in part because the deep sequestration of reduced organic carbon in sediments is one of the proposed mechanisms for the production of Earth's oxidized atmosphere[7]. However, a newly expanded isotopic database implicates carbonates in altered igneous oceanic crust (AOC) as an alternative source for the ^{13}C-depleted signal in many diamonds[8]. The stability of carbonated AOC at depth, until its partial melting in the deep asthenosphere and transition zone[9,10], reinforces the idea that the AOC could be the source of carbon in many superdeep diamonds[9,11]. However, thus far no geochemical signature has clearly related these diamonds to a carbonate-rich protolith. Additionally, given the expected carbon-depleted nature of slabs after slab melting in the transition zone, the source of carbon for lower-mantle diamonds remains unclear.

To evaluate the relative contributions of sediments, the AOC, and the convecting mantle to the deep-mantle carbon cycle, we analysed a suite of inclusions in diamonds for their oxygen isotope signature ($\delta^{18}O = (^{18}O/^{16}O_{sample})/(^{18}O/^{16}O_{VSMOW}) - 1$; VSMOW, Vienna Standard Mean Ocean Water), which is sensitive to the presence of recycled material.

[1]Canadian Centre for Isotopic Microanalysis, Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada. [2]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK. ✉e-mail: margoregier@gmail.com

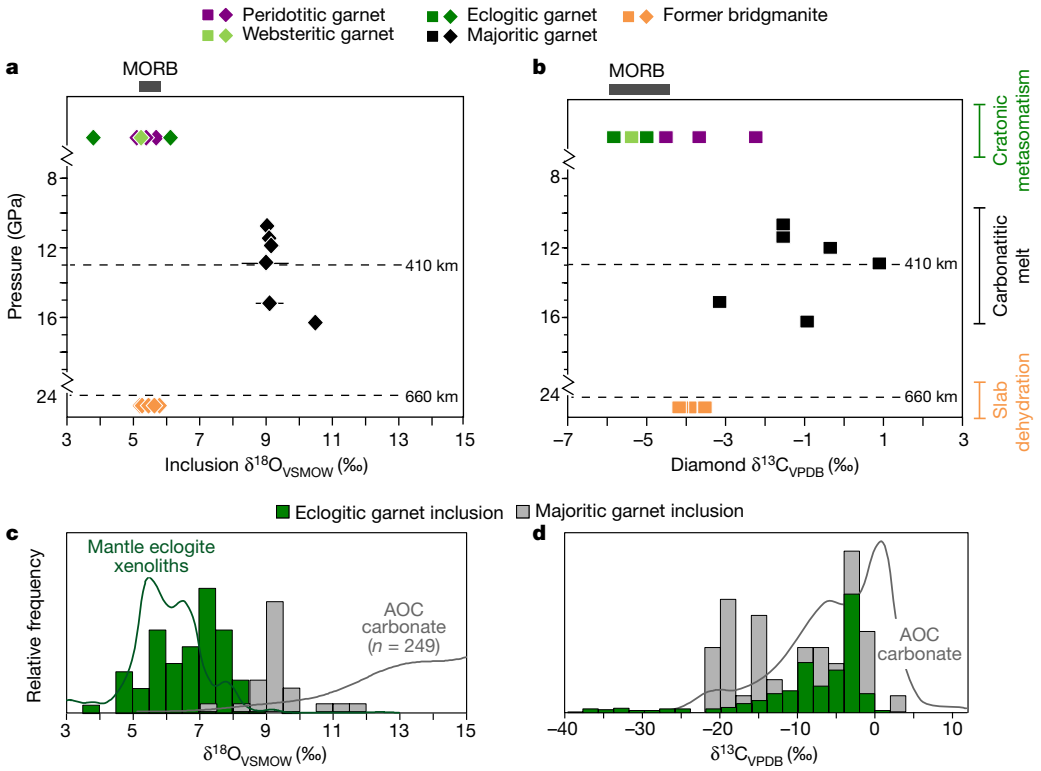


Fig. 1 | Stable-isotope compositions of diamonds and their mineral inclusions. a, **b**, Silicate inclusion $\delta^{18}O$ and diamond host $\delta^{13}C$ signatures[29,30] versus depth for a suite of Kankan diamonds. Errors are 2σ and often smaller than the symbol. MORB isotopic ranges are indicated by grey bars[16–18,41]. Inferred environments of formation are indicated on the right. **c**, A histogram of $\delta^{18}O$ for all measured majoritic garnet inclusions (grey, $n = 26$) has a more positive mode than that for eclogitic garnet inclusions (green, $n = 64$). Also plotted are probability density functions of eclogitic garnets from mantle xenoliths (green line; bandwidth of 0.2‰) and AOC carbonates (grey line; bandwidth of 1.9‰). **d**, Histograms of $\delta^{13}C$ for a worldwide database of eclogitic (green, $n = 467$) and majoritic garnet-bearing diamonds (grey, $n = 48$) and a probability density function (bandwidth of 1.17‰) for AOC carbonate. Note that the scale differs from that of Fig. 1b. References for the data shown in this figure are provided in the source data file.

Previous measurements of $\delta^{18}O$ in superdeep inclusions in diamond have been confined to two suites of asthenospheric and transition-zone inclusions from the Jagersfontein kimberlite (South Africa) and the Collier-4, Juina-5 and Machado alluvial deposits of the Juina region (Brazil)[12,13]. Here, we report $\delta^{18}O$ values of inclusions within a diamond suite from Kankan, Guinea that contains not only lithospheric and asthenospheric/transition-zone garnet inclusions, but also low-Al_2O_3 (<1.7 wt%) orthopyroxene (retrogressed bridgmanite) coexisting with ferropericlase, an assemblage from the uppermost lower mantle (~700 km)[14,15]. Kankan diamonds and their silicate inclusions are thus powerful probes of the carbon cycle from the lithosphere to lower mantle.

Lithospheric-to-transition-zone diamonds

Lithospheric garnet inclusions in Kankan diamonds can be divided into eclogitic and peridotitic suites based on major element chemistry[14]. The peridotitic suite $\delta^{18}O$ ranges from +5.3‰ ± 0.3‰ to +5.7‰ ± 0.2‰ (all uncertainities are 2σ), within error of the average mid-ocean-ridge basalt (MORB)-source mantle (+5.5‰)[16–18] and in equilibrium with cratonic peridotite olivine (+5.3‰ ± 0.2‰)[19]. The eclogitic suite has more varied $\delta^{18}O$ values (+3.8‰ to +6.1‰), indicative of an origin from altered oceanic crust (Fig. 1a, c). Given the required addition of unreasonably large amounts of oxygen to change mineral $\delta^{18}O$ values[20], these signatures are representative of the cratonic substrate, and not the introduced diamond-forming metasomatic agents.

Unlike lithospheric minerals, which can be definitively interpreted as having an eclogitic or peridotitic paragenesis using traditional major elemental classification schemes[21], sublithospheric majoritic garnets (characterized by an excess of Si) are more difficult to assign to a specific paragenesis[11]. Here, we categorize the majoritic garnets using an experimentally calibrated model, in which excess Si^{4+} in the majoritic endmember is charge-balanced with Na^{+} in eclogitic systems, or with divalent cations in Na-poor peridotitic compositions[11]. To quantify this scheme, we derive a parameter, Δperidotite, defined as the difference in divalent cations between the mineral inclusion and a purely meta-peridotitic majoritic garnet, normalized to the difference between the two endmembers. Thus, a majoritic garnet from a purely eclogitic system has a Δperidotite = 1, whereas one from a purely meta-peridoditic system has a Δperidotite = 0 (see Methods). The majority of majoritic garnet inclusions lie between these endmember trends ($0 < \Delta\text{peridotite} < 1$). This intermediate composition has been termed 'meta-pyroxenitic' (Fig. 2a)[11].

Because these meta-pyroxenitic inclusions are intermediate in major-element composition, we might also expect their $\delta^{18}O$ value to be intermediate compared to those reported for lithospheric eclogitic and peridotitic garnets. Instead, meta-pyroxenitic majoritic garnets from Kankan have much more extreme $\delta^{18}O$ values (+9.1‰ to +10.5‰) than the Kankan eclogitic garnet inclusions of lithospheric origin (+3.8‰ to +6.1‰). Similarly, meta-pyroxenitic majoritic garnets from Juina (Brazil)[13] and Jagersfontein (South Africa) diamonds[12] also have a considerably higher $\delta^{18}O$ mode than eclogitic garnet inclusions in diamonds worldwide (Fig. 1c), with all majoritic garnet values greater than +7.5‰ and some extending to even more extreme values (+12‰). Only 4% of data from a composite model of oceanic lithosphere approaches the average $\delta^{18}O$ of these majoritic garnets[22]. Even more striking, <0.05%

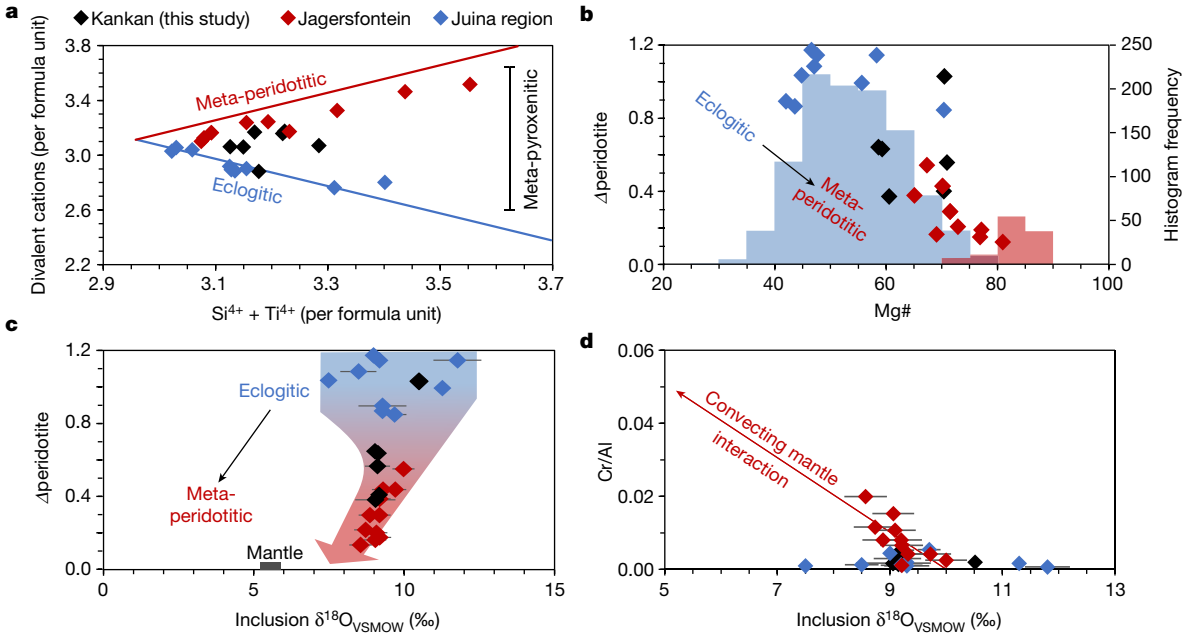


Fig. 2 | Elemental and isotopic composition of majoritic garnet inclusions. **a**, Divalent cations (Fe_{total}, Mg, Ca, Mn) versus Si and Ti per formula unit ([O] = 12) in majoritic garnets from Kankan, Jagersfontein and the Juina region. Red and blue lines show substitutions typical for meta-peridotitic and eclogitic compositions, respectively, which begin at the median value for eclogitic garnets (2.96, 3.11)[42]. **b**, Majoritic garnet Mg# [Mg/(Mg+Fe) × 100] versus Δperidotite, which indicates the deviation of an individual garnet from the meta-peridotitic substitution. The secondary-axis histograms show the distribution of Mg# in lithospheric eclogitic (blue) and peridotitic (red) garnet inclusions. **c**, $\delta^{18}O$ versus Δperidotite for majoritic garnet. **d**, Cr/Al and $\delta^{18}O$ in majoritic garnets. The red line is a linear regression ($r^2 = 0.6$) for the Jagersfontein data, which trends from low Cr/Al eclogitic majorites to more meta-peridotitic, high-Cr/Al majorites with lower $\delta^{18}O$. Error bars in **c**, **d** are 2σ and may be smaller than the symbols. References for the data shown in this figure are provided in the source data file.

of oceanic lithosphere bulk rocks extend to the +12‰ observed in some majoritic garnets[22]. Hence, we conclude that there must be a unique source for the highly elevated $\delta^{18}O$ in meta-pyroxenitic superdeep inclusions.

The oxygen isotope compositions of asthenosphere-to-transition-zone majoritic garnets clearly require a crustal input, which must be sourced somewhere in the subducting oceanic slab. Two potential reservoirs of high $\delta^{18}O$ are the carbon-bearing constituents of sediments on top of the oceanic crust and carbonate in the AOC itself. Sediments are generally dominated by positive $\delta^{13}C$ 'marine carbonate' ($\delta^{13}C = (^{13}C/^{12}C_{sample})/(^{13}C/^{12}C_{VPDB}) - 1$; VPDB, Vienna Pee Dee Belemnite), but may be locally ^{13}C-depleted owing to the presence of 'reduced organic-rich carbon', which is comprised of organic carbon from marine and terrestrial organisms living near continental margins[1]. In comparison, much of the carbon present in the AOC is ^{13}C-enriched carbonate that precipitated in equilibrium with dissolved inorganic carbon (DIC), called 'normal' or 'DIC-equilibrium carbonate'[8]. However, recent studies have documented that AOC also includes ^{13}C-depleted carbonate precipitated from biologically or kinetically fractionated DIC[8]; we label this endmember 'biogenic carbonate'.

The recent identification of ^{13}C-depleted biogenic carbonate in the AOC[8] challenges the common assumption that ^{13}C-depleted diamonds invariably originate from deeply subducted sediment[4,6]. To investigate the source of the ^{13}C-depleted signal further, we examine the worldwide database of $\delta^{15}N$ in diamond ($\delta^{15}N = (^{15}N/^{14}N_{sample})/(^{14}N/^{13}N_{air}) - 1$), because this isotopic system can more clearly discriminate between AOC and sediment sources. We find that ~20% of all eclogitic diamonds of lithospheric origin have lower $\delta^{15}N$ (<−7‰) than the convecting mantle, and ~80% have $\delta^{15}N < 0$, suggesting that a portion of the subducted endmember must have strongly negative $\delta^{15}N$ (Fig. 3). Organic-rich sediments cannot satisfy this requirement, because their $\delta^{15}N$ values are almost exclusively positive[23]. By contrast, the AOC can satisfy this condition because it spans a large range of $\delta^{15}N$ (−12‰ to +12‰),

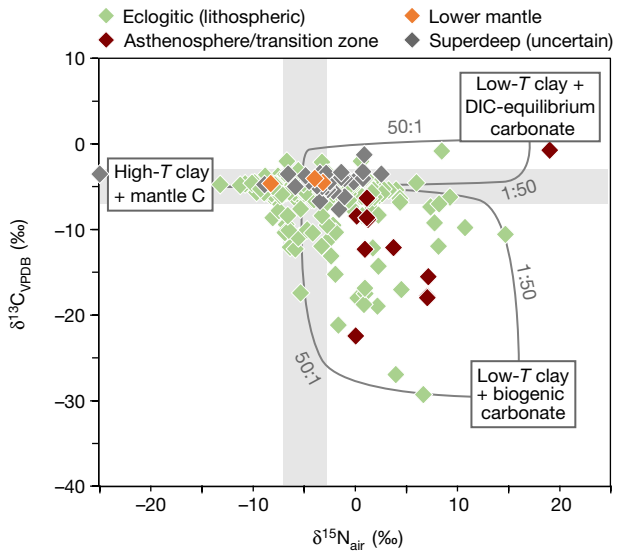


Fig. 3 | Worldwide database of $\delta^{13}C$ and $\delta^{15}N$ for diamonds of lithospheric and superdeep origin. MORB isotopic signatures are represented by grey bands[16–18,41]. Bracketing the $\delta^{13}C$ and $\delta^{15}N$ data for diamonds are three AOC endmembers composed of nitrogen-bearing high-temperature (T) clay or low-T clay, as well as carbon-bearing convecting mantle, DIC-equilibrium carbonate or biogenic carbonate. The mixing lines between the endmembers and the convecting mantle are $(N/C)_{mantle}/(N/C)_{AOC}$ = 50:1 and 1:50 (ref. [8]). Superdeep diamonds include those of asthenospheric and transition-zone origin (majoritic garnet and Ca–silicate inclusions), lower-mantle origin (ferropericlase + $MgSiO_3$ or $CaSiO_3$ inclusions) and diamonds of uncertain superdeep origin (individual ferropericlase, ilmenite, coesite and chromite inclusions). We note that asthenospheric/transition-zone diamonds have exclusively positive $\delta^{15}N$, but variable $\delta^{13}C$. The 2σ error bars are smaller than symbols. References for the data shown in this figure are provided in the source data file.

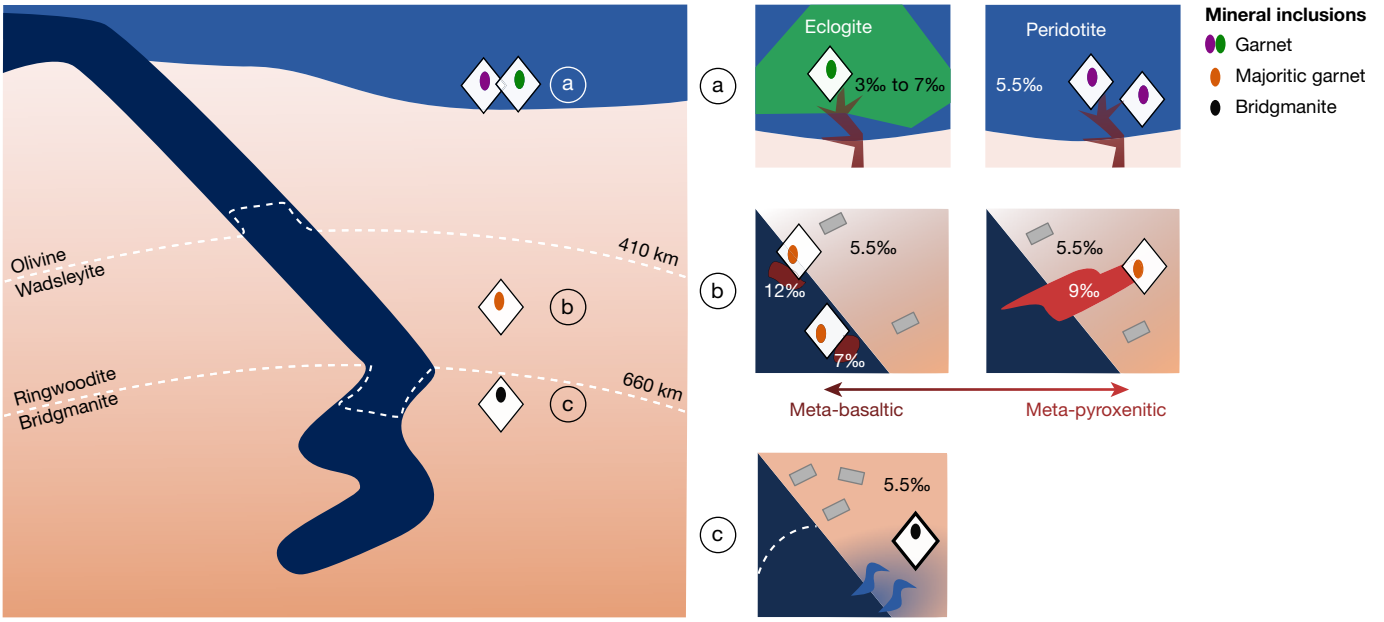


Fig. 4 | Model of diamond formation in the lithosphere, transition zone and lower mantle. a, Lithospheric diamond forms by fluid or melt metasomatism of eclogitic and peridotitic substrates[43], but the $\delta^{18}O$ value of the inclusions is buffered by the host lithology[20]. **b**, In the transition zone, the carbonate-rich upper portion of a subducting slab produces carbonatitic melt. Diamonds and majoritic garnet inclusions crystallize during the interaction of the carbonatitic melt with reduced, metal-bearing convecting mantle. The short melt migration path and the limited interaction with convecting mantle produces majoritic garnet with eclogitic compositions and elevated $\delta^{18}O$, directly reflecting the local carbonated AOC melt source. Greater levels of interaction with the convecting mantle are reflected in the lower $\delta^{18}O$ and increasingly ultrabasic, meta-pyroxenitic character of some majorites. **c**, As the slab penetrates into the lower mantle, the negative pressure–temperature slope (the Clapeyron slope) of the post-spinel transition[44] and the delayed garnet-to-perovskite transition in metabasaltic lithologies[45] retards the formation of lower-mantle minerals (dotted white line). The transition to a lower-mantle mineralogy leads to slab dehydration and the hydration of the surrounding mantle. The hydrated ambient mantle releases carbon from its metallic iron hosts to form diamond.

reflecting ^{15}N depletion in high-temperature clays and ^{15}N enrichment in low-temperature clays[23,24]. Thus, following Li et al.[8], we suggest that the isotopic variability defining most eclogitic diamonds from the lithosphere can be modelled by mixing between three AOC endmembers: (1) nitrogen-bearing high-temperature clay with mantle-derived carbon, (2) nitrogen-bearing low-temperature clay with DIC-equilibrium carbonate and (3) nitrogen-bearing low-temperature clay with biogenic carbonate (Fig. 3). Given the absence of a strong sedimentary $\delta^{15}N$ signal in lithospheric eclogitic diamonds, we suggest that sediments are an even more unlikely source of carbon in the deeper mantle sampled by sublithospheric diamonds. Interestingly, all reported asthenospheric and transition-zone diamonds have $\delta^{15}N > 0$ (Fig. 3), suggesting that diamond formation in that section of the mantle is driven strictly by the uppermost portions of the AOC that are rich in carbonate, that is, endmembers 2 and 3.

A carbonate-rich oceanic-crust origin for asthenospheric-to-transition-zone diamonds is consistent with the majoritic garnet inclusions studied here. Not only are the inclusions offset towards the higher $\delta^{18}O$ values recorded in oceanic-crust carbonates (Fig. 1c), but the formation depths of majoritic garnets (7–19 GPa; Extended Data Fig. 1) also correlate well with the experimentally constrained pressures at which subducted, carbonated metabasalt may melt (11–21 GPa)[25]. Experiments have demonstrated that slab-derived carbonatitic melt will crystallize diamond after injection into and reaction with the surrounding reduced convecting mantle, and will crystallize majoritic garnet upon cooling below the liquidus temperature[9,10,26]. We propose that the isotopic characteristics of these inclusions are principally derived from the AOC carbonate component, whereas the elemental characteristics depend on the degree of interaction with the surrounding mantle. High-Δperidotite (more eclogitic) majoritic garnets probably crystallized from Mg-poor, slab-derived carbonatitic melts ($Mg/O \approx 0.05$; Mg number, $Mg\# \approx 55$)[9] that experienced little mantle contamination before carbonate reduction and the resulting solidification and precipitation of diamond[27]. Greater extents of interaction with the surrounding Mg-rich convecting mantle ($Mg/O \approx 0.34$; $Mg\# \approx 90$)[28] is evident in those meta-pyroxenitic majoritic garnets with increased Mg#, but only extreme levels of interaction between melt and mantle could have produced the shift to slightly lower $\delta^{18}O$ that is seen in Jagersfontein majoritic garnets (Extended Data Fig. 2). Thus, the reaction between AOC-derived carbonatitic melt and convecting mantle created intermediate 'mixed' elemental compositions, yet maintained the high $\delta^{18}O$ values that are characteristic of extremely ^{18}O-enriched carbonated AOC[8].

Lower-mantle diamonds

Whereas the strong ^{18}O enrichment in majoritic garnets is related to subducted carbonated crust in the asthenosphere and transition zone, the first $\delta^{18}O$ measurements made here of lower-mantle retrogressed bridgmanites show no such ^{18}O enrichment. Instead, the enstatite $\delta^{18}O$ (+5.3‰ to +5.8‰), the average Mg# of Kankan bridgmanite and ferropericlase inclusions (95.0 and 86.7, respectively), and the host-diamond $\delta^{13}C$ (−3.5‰ to −4.1‰)[29,30] are all similar to estimates of fertile mantle that has not experienced substantial exchange with recycled crustal material (Extended Data Table 1)[31–33]. This lack of an obvious crustal signature in sublithospheric diamonds and their inclusions is unusual[34], probably because slab-derived carbonate is generally required to increase carbon concentrations and stabilize diamond in the metal-bearing reduced deep mantle (>8 GPa)[2,35,36]. Given that the lower mantle is estimated to have 1 wt% metal[35] and 16–500 ppm C (ref. [37]), iron metal and/or iron carbides are the dominant carbon-bearing phases at these depths[2,36]. In order to produce macrocrystalline lower

mantle diamonds without carbonate input, the carbon in metal alloys needs to be mobilized and locally concentrated. One way of achieving this is via the introduction of a dehydrating slab into the lower mantle[38,39], because carbon-bearing metal alloys are unstable in hydrated environments[40]. The local destabilization of ~1 wt% carbon-bearing metal alloys requires limited lithosphere-derived H_2O (ref. [40]), which would not markedly affect the ambient-lower-mantle $\delta^{18}O$. Therefore, we speculate that dehydration of a carbonate-depleted subducting oceanic lithosphere can trigger the metasomatic mobilization of ambient carbon for lower-mantle diamond formation, without imposing a crustal signature on the resulting diamonds and their inclusions.

The lithospheric-to-lower-mantle carbon cycle

The contrasting stable-isotope compositions of diamonds and their silicate inclusions at lithospheric, transition-zone and lower-mantle depths suggest profound differences in the modes of diamond formation and in the behaviour of volatiles through these mantle regions (Fig. 4). The absence of a clear sediment-derived geochemical signal at diamond-forming depths has implications for the efficiency with which carbon is recycled within Earth's mantle, and suggests that volatile elements in sediments may be efficiently recycled back to the surface during arc volcanism or stored in shallow accretionary prisms[5]. Instead, we document geochemical evidence for the deep cycling of carbonated AOC as a source of lithospheric-to-transition-zone diamonds. Furthermore, diamonds from even deeper, in the uppermost lower mantle show no evidence of a subducted crustal carbon or oxygen mass flux. We suggest that these diamonds crystallized after the dehydrating slab triggered the mobilization of convecting-mantle carbon from its metallic hosts. This change of the diamond-forming environment from a carbonated slab melt in the transition zone to a slab-hydrated lower mantle is consistent with experimental evidence that demonstrates major obstacles to transporting AOC carbonates to lower-mantle depths along typical slab thermal trajectories[9]. Our study, therefore, supports a barrier to carbon subduction above the lower mantle[9].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2676-z.

1. Plank, T. & Manning, C. E. Subducting carbon. *Nature* **574**, 343–352 (2019).
2. Dasgupta, R. & Hirschmann, M. M. The deep carbon cycle and melting in Earth's interior. *Earth Planet. Sci. Lett.* **298**, 1–13 (2010).
3. Creager, K. C. & Jordan, T. H. Slab penetration into the lower mantle. *J. Geophys. Res.* **89**, 3031–3049 (1984).
4. Cartigny, P. Stable isotopes and the origin of diamond. *Elements* **1**, 79–84 (2005).
5. Kelemen, P. B. & Manning, C. E. Reevaluating carbon fluxes in subduction zones, what goes down, mostly comes up. *Proc. Natl Acad. Sci. USA* **112**, 3997–4006 (2015).
6. Sobolev, V. S. & Sobolev, N. V. New proof of very deep subsidence of eclogitized crustal rocks. *Dokl. Acad. Nauk SSSR*. **250**, 88–90 (1982).
7. Duncan, M. S. & Dasgupta, R. Rise of Earth's atmospheric oxygen controlled by efficient subduction of organic carbon. *Nat. Geosci*. **10**, 387–392 (2017).
8. Li, K., Li, L., Pearson, D. G. & Stachel, T. Diamond isotope compositions indicate altered igneous oceanic crust dominates deep carbon recycling. *Earth Planet. Sci. Lett*. **516**, 190–201 (2019).
9. Thomson, A. R., Walter, M. J., Kohn, S. C. & Brooker, R. A. Slab melting as a barrier to deep carbon subduction. *Nature* **529**, 76–79 (2016).
10. Walter, M. J. et al. Primary carbonatite melt from deeply subducted oceanic crust. *Nature* **454**, 622–625 (2008).
11. Kiseeva, E. S. et al. Metapyroxenite in the mantle transition zone revealed from majorite inclusions in diamonds. *Geology* **41**, 883–886 (2013).
12. Ickert, R. B., Stachel, T., Stern, R. A. & Harris, J. W. Extreme 18O-enrichment in majorite constrains a crustal origin of transition zone diamonds. *Geochem. Perspect. Lett*. **1**, 65–74 (2015).
13. Burnham, A. D. et al. Stable isotope evidence for crustal recycling as recorded by superdeep diamonds. *Earth Planet. Sci. Lett*. **432**, 374–380 (2015).
14. Stachel, T., Brey, G. P. & Harris, J. W. Kankan diamonds (Guinea) I: from the lithosphere down to the transition zone. *Contrib. Mineral. Petrol*. **140**, 1–15 (2000).
15. Stachel, T., Harris, J. W., Brey, G. P. & Joswig, W. Kankan diamonds (Guinea) II: lower mantle inclusion parageneses. *Contrib. Mineral. Petrol*. **140**, 16–27 (2000).
16. Cooper, K. M., Eiler, J. M., Sims, K. W. W. & Langmuir, C. H. Distribution of recycled crust within the upper mantle: insights from the oxygen isotope composition of MORB from the Australian–Antarctic Discordance. *Geochem. Geophys. Geosyst*. **10**, Q12004 (2009).
17. Cooper, K. M., Eiler, J. M., Asimow, P. D. & Langmuir, C. H. Oxygen isotope evidence for the origin of enriched mantle beneath the mid-Atlantic ridge. *Earth Planet. Sci. Lett*. **220**, 297–316 (2004).
18. Eiler, J. M., Schiano, P., Kitchen, N. & Stolper, E. M. Oxygen-isotope evidence for recycled crust in the sources of mid-ocean-ridge basalts. *Nature* **403**, 530–534 (2000).
19. Regier, M. E. et al. An oxygen isotope test for the origin of Archean mantle roots. *Geochem. Perspect. Lett*. **9**, 6–10 (2018).
20. Riches, A. J. V. et al. In situ oxygen-isotope, major-, and trace-element constraints on the metasomatic modification and crustal origin of a diamondiferous eclogite from Roberts Victor, Kaapvaal Craton. *Geochim. Cosmochim. Acta* **174**, 345–359 (2016).
21. Grütter, H. S., Gurney, J. J., Menzies, A. H. & Winter, F. An updated classification scheme for mantle-derived garnet, for use by diamond explorers. *Lithos* **77**, 841–857 (2004).
22. Ickert, R. B., Stachel, T., Stern, R. A. & Harris, J. W. Diamond from recycled crustal carbon documented by coupled $\delta^{18}O$–$\delta^{13}C$ measurements of diamonds and their inclusions. *Earth Planet. Sci. Lett*. **364**, 85–97 (2013).
23. Li, L., Zheng, Y. F., Cartigny, P. & Li, J. Anomalous nitrogen isotopes in ultrahigh-pressure metamorphic rocks from the Sulu orogenic belt: effect of abiotic nitrogen reduction during fluid-rock interaction. *Earth Planet. Sci. Lett*. **403**, 67–78 (2014).
24. Li, L., Bebout, G. E. & Idleman, B. D. Nitrogen concentration and $\delta^{15}N$ of altered oceanic crust obtained on ODP Legs 129 and 185: insights into alteration-related nitrogen enrichment and the nitrogen subduction budget. *Geochim. Cosmochim. Acta* **71**, 2344–2360 (2007).
25. Kiseeva, E. S., Litasov, K. D., Yaxley, G. M., Ohtani, E. & Kamenetsky, V. S. Melting and phase relations of carbonated eclogite at 9–21 GPa and the petrogenesis of alkali-rich melts in the deep mantle. *J. Petrol*. **54**, 1555–1583 (2013).
26. Bobrov, A. V., Litvin, Y. A., Bindi, L. & Dymshits, A. M. Phase relations and formation of sodium-rich majoritic garnet in the system $Mg_3Al_2Si_3O_{12}$–$Na_2MgSi_5O_{12}$ at 7.0 and 8.5 GPa. *Contrib. Mineral. Petrol*. **156**, 243–257 (2008).
27. Rohrbach, A. & Schmidt, M. W. Redox freezing and melting in the Earth's deep mantle resulting from carbon–iron redox coupling. *Nature* **472**, 209–212 (2011).
28. Ringwood, A. E. *Composition and Petrology of the Earth's Mantle* (McGraw-Hill, 1975).
29. Stachel, T., Harris, J. W., Aulbach, S. & Deines, P. Kankan diamonds (Guinea) III: $\delta^{13}C$ and nitrogen characteristics of deep diamonds. *Contrib. Mineral. Petrol*. **142**, 465–475 (2002).
30. Palot, M., Pearson, D. G., Stern, R. A., Stachel, T. & Harris, J. W. Isotopic constraints on the nature and circulation of deep mantle C–H–O–N fluids: carbon and nitrogen systematics within ultra-deep diamonds from Kankan (Guinea). *Geochim. Cosmochim. Acta* **139**, 26–46 (2014).
31. Cartigny, P., Palot, M., Thomassot, E. & Harris, J. W. Diamond formation: a stable isotope perspective. *Annu. Rev. Earth Planet. Sci*. **42**, 699–732 (2014).
32. Katsura, T. & Ito, E. Determination of Fe–Mg partitioning between perovskite and magnesiowüstite. *Geophys. Res. Lett*. **23**, 2005–2008 (1996).
33. Wood, B. J. Phase transformations and partitioning relations in peridotite under lower mantle conditions. *Earth Planet. Sci. Lett*. **174**, 341–354 (2000).
34. Stachel, T. Diamonds from the asthenosphere and the transiton zone. *Eur. J. Mineral*. **13**, 883–892 (2001).
35. Frost, D. J. & McCammon, C. A. The redox state of Earth's mantle. *Annu. Rev. Earth Planet. Sci*. **36**, 389–420 (2008).
36. Rohrbach, A., Ghosh, S., Schmidt, M. W., Wijbrans, C. H. & Klemme, S. The stability of Fe–Ni carbides in the Earth's mantle: evidence for a low Fe–Ni–C melt fraction in the deep mantle. *Earth Planet. Sci. Lett*. **388**, 211–221 (2014).
37. Hirschmann, M. M. & Dasgupta, R. The H/C ratios of Earth's near-surface and deep reservoirs, and consequences for deep Earth volatile cycles. *Chem. Geol*. **262**, 4–16 (2009).
38. Schmandt, B., Jacobsen, S. D., Becker, T. W., Liu, Z. & Dueker, K. G. Dehydration melting at the top of the lower mantle. *Science* **344**, 1265–1268 (2014).
39. Faccenda, M. Water in the slab: a trilogy. *Tectonophysics* **614**, 1–30 (2014).
40. Zhu, F., Li, J., Liu, J., Dong, J. & Liu, Z. Metallic iron limits silicate hydration in Earth's transition zone. *Proc. Natl Acad. Sci. USA* **116**, 22526–22530 (2019).
41. Marty, B. & Zimmermann, L. Volatiles (He, C, N, Ar) in mid-ocean ridge basalts: assesment of shallow-level fractionation and characterization of source composition. *Geochim. Cosmochim. Acta* **63**, 3619–3633 (1999).
42. Beard, B. L. et al. Petrography and geochemistry of eclogites from the Mir kimberlite, Yakutia, Russia. *Contrib. Mineral. Petrol*. **125**, 293–310 (1996).
43. Stachel, T. & Luth, R. W. Diamond formation — where, when and how? *Lithos* **220–203**, 200–220 (2015).
44. Irifune, T. & Ringwood, A. E. Phase transformations in a harzburgite composition to 26 GPa: implications for dynamical behaviour of the subducting slab. *Earth Planet. Sci. Lett*. **86**, 365–376 (1987).
45. Ono, S., Ito, E. & Katsura, T. Mineralogy of subducted basaltic crust (MORB) from 25 to 37 GPa, and chemical heterogeneity of the lower mantle. *Earth Planet. Sci. Lett*. **190**, 57–63 (2001).

Methods

Enstatite and garnet inclusions in diamond were analysed using a Cameca IMS 1280 multicollector ion microprobe with ~2 nA $^{33}Cs^+$ primary beam and 20 keV impact energy. The analytical methods and standards used for garnets have been published previously[46]. The presence of high-Cr_2O_3 lithospheric garnets required the development of a new matrix correction. Olivine and high-Cr_2O_3 garnet pairs from depleted peridotite xenoliths were cast into epoxy and pressed into indium mounts along with garnet and olivine reference material. Olivine $\delta^{18}O$ values were all within the error of the mantle values, suggesting that the associated garnets should be also within the expected convecting-mantle $\delta^{18}O$ range. Instead, a plot of garnet Cr_2O_3 versus $\delta^{18}O$ defines a positive slope that reaches ~1‰ above the mantle range at high Cr_2O_3 contents (Extended Data Fig. 3). A Cr-related matrix effect has been previously suggested[47], but variable laser fluorination yields of Cr-rich minerals have inhibited a robust determination of the calibration. Our method bypasses the need for laser fluorination of high-Cr garnets, being instead based on laser fluorination of a low-Cr garnet (S0068) and a reasonable assumption of mineral isotopic equilibrium at mantle temperatures. Using this calibration, the 95% confidence uncertainty estimates for $\delta^{18}O_{VSMOW}$ for garnets averaged ±0.29‰. Enstatite $\delta^{18}O$ measurements also required the development of a new calibration for Mg# (Extended Data Fig. 4) using laser fluorination results[48] for sample F866 (Mg# = 94.1) and CCIM standard S0170 (Mg# = 91.2). For the analyses of unknown enstatites, the 95% confidence uncertainty estimates for $\delta^{18}O_{VSMOW}$ average ±0.21‰. Adjacent to each ion probe crater, major-element data were collected on a Cameca SX100 electron probe microanalyser with five wavelength-dispersive spectrometers at 20 keV energy and 20 nA beam current at 1 µm diameter. The counting time was 30 s for all elements. The detection limits are available at the bottom of Supplementary Table 1 and standards are reported in Extended Data Table 2.

The parameter Δperidotite was applied to majoritic garnets to describe the deviation from a pure meta-peridotitic endmember:

$$\Delta\text{peridotite} = \frac{|(\text{Mg} + \text{Ca} + \text{Fe} + \text{Mn})_s - [m_p(\text{Si} + \text{Ti})_s + b_p]|}{|[m_e(\text{Si} + \text{Ti})_s + b_e] - [m_p(\text{Si} + \text{Ti})_s + b_p]|} \quad (1)$$

where m is the slope of the endmember substitution, b is the y intercept of the endmember, p and e denote the meta-peridotite and eclogitic endmembers, respectively, and s represents the majoritic garnet sample. Fe indicates total iron.

Data availability

Geochemical data that support the findings of this study are available at https://ecl.earthchem.org/view.php?id=1580. Source data are provided with this paper.

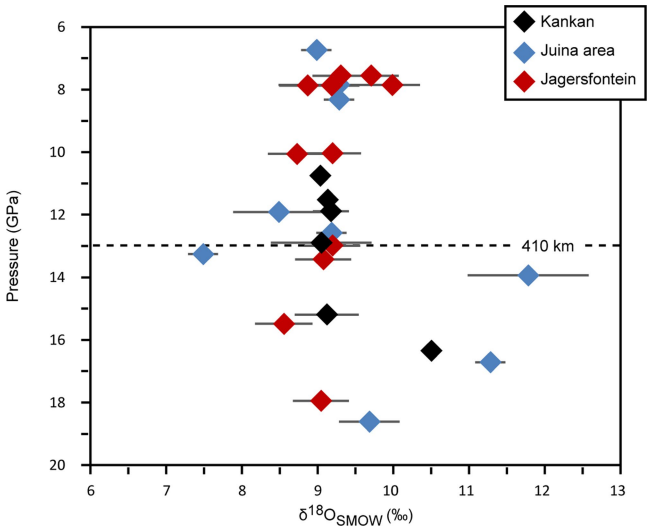
46. Ickert, R. B. & Stern, R. A. Matrix corrections and error analysis in high-precision SIMS $^{18}O/^{16}O$ measurements of Ca–Mg–Fe garnet. *Geostand. Geoanal. Res.* **37**, 429–448 (2013).
47. Wang, Z., Bucholz, C., Skinner, B., Shimizu, N. & Eiler, J. Oxygen isotope constraints on the origin of high-Cr garnets from kimberlites. *Earth Planet. Sci. Lett.* **312**, 337–347 (2011).
48. Mattey, D., Lowry, D. & Macpherson, C. Oxygen isotope composition of mantle peridotite. *Earth Planet. Sci. Lett.* **128**, 231–241 (1994).
49. Beyer, C. & Frost, D. J. The depth of sub-lithospheric diamond formation and the redistribution of carbon in the deep mantle. *Earth Planet. Sci. Lett.* **461**, 30–39 (2017).
50. McDonough, W. F. & Rudnick, R. L. Mineralogy and composition of the upper mantle. *Rev. Mineral. Geochem.* **37**, 139–164 (1998).
51. Zheng, Y.-F. Calculation of oxygen isotope fractionation in anhydrous silicate minerals. *Geochim. Cosmochim. Acta* **57**, 1079–1091 (1993); erratum **57**, 3199 (1993).
52. Lowry, D., Mattey, D. P. & Harris, J. W. Oxygen isotope composition of syngenetic inclusions in diamond from the Finsch Mine, RSA. *Geochim. Cosmochim. Acta* **63**, 1825–1836 (1999).

Acknowledgements We acknowledge Canada Excellence Research Chairs and the Deep Carbon Observatory for funding this study. We thank Diamond Trading Company (a member of the DeBeers Group of Companies) for the donation to J.W.H. of the diamonds used in this study.

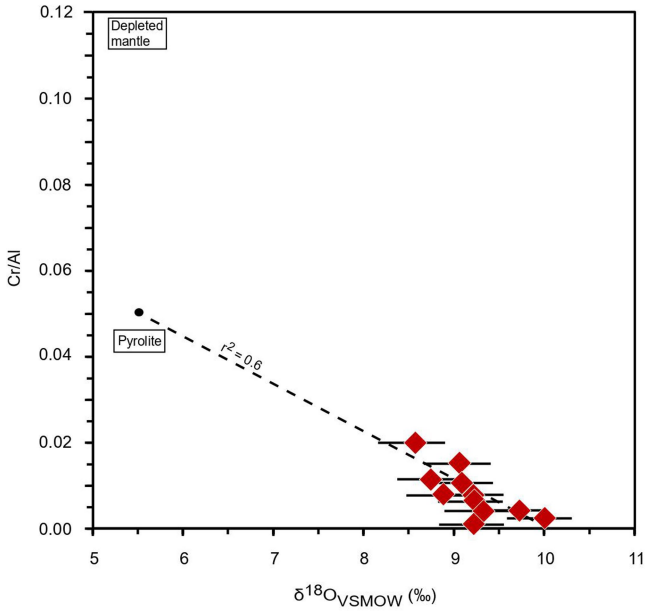
Author contributions M.E.R. and R.A.S. collected the data. M.E.R. provided the initial data interpretation and manuscript. Input from all other authors improved the interpretation and writing. J.W.H. provided the samples.

Competing interests The authors declare no competing interests.

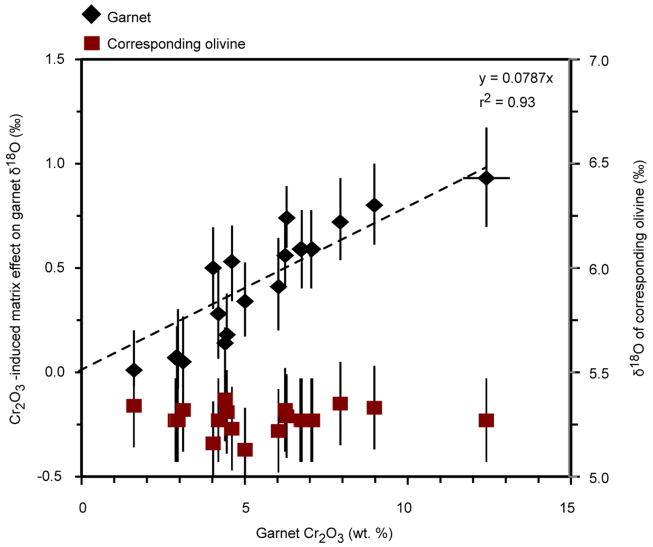
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2676-z.
Correspondence and requests for materials should be addressed to M.E.R.
Peer review information *Nature* thanks John Eiler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



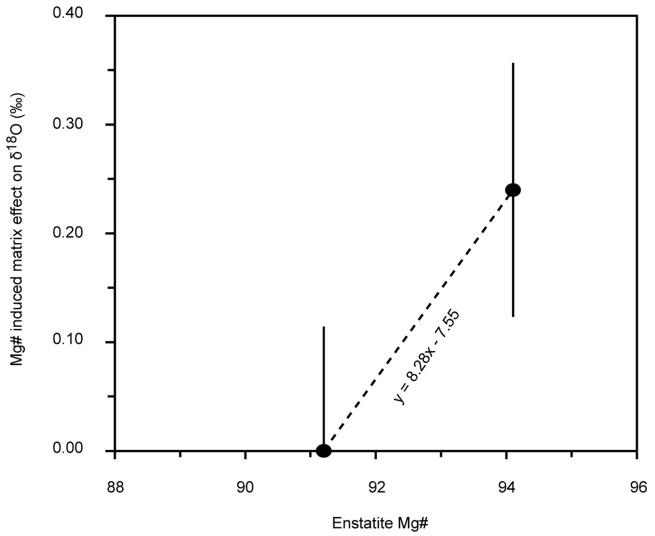
Extended Data Fig. 1 | Oxygen isotope values for majoritic garnet inclusions versus pressure of formation. Majoritic garnet inclusions include those from the Juina area (Brazil), Jagersfontein (South Africa) and Kankan (Guinea) majorites. Oxygen isotope values are shown versus pressure estimates[49]. Error bars are 2σ (refs. [12,13]).



Extended Data Fig. 2 | Oxygen isotope values versus Cr/Al for Jagersfontein majoritic garnets. A linear regression (r^2 = 0.6) intersects a 5.5‰ mantle assimilate with a Cr/Al content of ~0.05, whereas primitive mantle has a Cr/Al of ~0.04 (ref. [28]) and mildly depleted mantle has a Cr/Al of ~0.11 (ref. [50]). Error bars are 2σ.



Extended Data Fig. 3 | Ion-probe $\delta^{18}O$ calibration for Cr-rich garnets. The oxygen isotopic composition of coexisting garnets and olivines from peridotitic mantle xenoliths were analysed using the ion probe to determine the instrumental fractionation associated with the Cr_2O_3 content of garnets. The plot defines the olivine $\delta^{18}O$ and the deviation of the measured garnet $\delta^{18}O$ from equilibrium after Ca# matrix correction[46], versus the Cr_2O_3 contents of the garnets. Because all the olivines have $\delta^{18}O$ within the error of the mantle, we assume isotopic equilibrium between garnet and olivine[51] and contend that the trend of SIMS-determined garnet $\delta^{18}O$ with Cr_2O_3 content is a matrix effect. The trendline indicates the correction of the $\delta^{18}O$ values to a hypothetical Cr-free garnet. Errors are 2σ.



Extended Data Fig. 4 | Ion-probe $\delta^{18}O$ calibration for enstatite Mg#. The instrumental mass fractionation with enstatite Mg# was assessed using reference material S0170 (Mg# of 91.2; laser fluorination $\delta^{18}O$ of +5.64‰) and S0444 (Mg# of 94.1; laser fluorination $\delta^{18}O$ of +5.76)[52]. Error bars incorporate 0.10‰ analytical uncertainty in the laser fluorination measurements.

Extended Data Table 1 | Mg# of bridgmanite and ferropericlase in experiments and natural inclusions in diamond

	Fo91 starting material (ref. [32])	Fo89 starting material (ref. [32])	Kankan Mg# average	Kankan Mg# 2σ
Bridgmanite	95.5	94.7	95.0	2.5
Ferropericlase	87.0	86.3	86.7	1.4

The Mg# values of Kankan bridgmanite and ferropericlase fall between those produced in Fo91 and Fo89 experiments[32] and overlap with mineralogical Mg# estimates (91–93 and 84–87 for bridgmanite and ferropericlase, respectively) for pyrolite with a Mg# of 90 (ref. [33]). This suggests that the lower-mantle Kankan inclusions are derived from a composition similar to that of pyrolite with a Mg# of ~90 (ref. [28]).

Extended Data Table 2 | Standards used for electron probe microanalyser analyses

Element	Analyzing crystal	Standard (inclusion mineralogy)
Si kα	LTAP	Fo90.5 olivine from Harvard (olivine) Enstatite (enstatite) Frank Smith pyrope garnet (garnet)
Ti kα	PET	Rutile from MTI
Al kα	TAP	Frank Smith pyrope garnet
Cr kα	PET	Chromium oxide from Alfa Aesar
Fe kα	LLIF	Fe_2SiO_4 fayalite from Rockport, MA
Ni kα	LLIF	Nickel from Alfa Aesar
Mn kα	LLIF	Spessartine from Navegadora Mine, Brazil
Mg kα	LTAP	Fo90.5 olivine from Harvard (olivine) Enstatite (enstatite) Frank Smith pyrope garnet (garnet)
Ca kα	LPET	USNM 115900 labradorite from Oregon
Na kα	TAP	Harvard 131705 albite from Virginia
K kα	LPET	Sanidine from Itrongay, Madagascar

Secondary standards included the Gore garnet from New York, Fo90 from San Carlos and enstatite H131709 from the Harvard collection.

Article

Origin and elaboration of a major evolutionary transition in individuality

https://doi.org/10.1038/s41586-020-2653-6

Received: 2 October 2018

Accepted: 3 June 2020

Published online: 2 September 2020

Check for updates

Ab. Matteen Rafiqi[1,2,3], Arjuna Rajakumar[1,3] & Ehab Abouheif[1]✉

Obligate endosymbiosis, in which distantly related species integrate to form a single replicating individual, represents a major evolutionary transition in individuality[1–3]. Although such transitions are thought to increase biological complexity[1,2,4–6], the evolutionary and developmental steps that lead to integration remain poorly understood. Here we show that obligate endosymbiosis between the bacteria *Blochmannia* and the hyperdiverse ant tribe Camponotini[7–11] originated and also elaborated through radical alterations in embryonic development, as compared to other insects. The Hox genes *Abdominal A* (*abdA*) and *Ultrabithorax* (*Ubx*)—which, in arthropods, normally function to differentiate abdominal and thoracic segments after they form—were rewired to also regulate germline genes early in development. Consequently, the mRNAs and proteins of these Hox genes are expressed maternally and colocalize at a subcellular level with those of germline genes in the germplasm and three novel locations in the freshly laid egg. *Blochmannia* bacteria then selectively regulate these mRNAs and proteins to make each of these four locations functionally distinct, creating a system of coordinates in the embryo in which each location performs a different function to integrate *Blochmannia* into the Camponotini. Finally, we show that the capacity to localize mRNAs and proteins to new locations in the embryo evolved before obligate endosymbiosis and was subsequently co-opted by *Blochmannia* and Camponotini. This pre-existing molecular capacity converged with a pre-existing ecological mutualism[12,13] to facilitate both the horizontal transfer[10] and developmental integration of *Blochmannia* into Camponotini. Therefore, the convergence of pre-existing molecular capacities and ecological interactions—as well as the rewiring of highly conserved gene networks—may be a general feature that facilitates the origin and elaboration of major transitions in individuality.

The obligate endosymbiosis between the bacteria *Blochmannia* and ants of the Camponotini is thought to have contributed to the ecological and evolutionary success of these organisms[10,11,14–21]. Phylogenetic evidence suggests that the ancestor of *Blochmannia* was horizontally transferred from hemipteran bugs (a distantly related order of insects) to the most recent common ancestor of the Camponotini approximately 51 million years ago[10,16,22]. *Blochmannia* enhances nutrition by increasing amino acid synthesis, which can regulate the size distribution of worker ants[11,14,16,23]. Ants, in turn, provide *Blochmannia* with a protected cellular environment for proliferation and ensure the strict vertical transmission of these bacteria through the germline[17–19,24]. As a consequence, *Blochmannia* and Camponotini have co-evolved and their phylogenies are congruent[20,25,26].

Embryogenesis is radically altered

In ants, wasps and flies, the germplasm is a maternally inherited region of cytoplasm that is localized to the posterior pole of oocytes and freshly laid eggs, where it has a dual function in specifying the germline and the embryonic posterior[27–31]. The mRNAs and/or proteins of a group of highly conserved 'germline genes' are localized together in the germplasm[27] (Supplementary Table 1). To investigate whether the integration of *Blochmannia* into Camponotini influences the germplasm, we first determined the localization of mRNAs or proteins of germline genes in the freshly laid eggs of *Lasius niger*, an early-branching species that is in the same subfamily as the Camponotini (Formicinae) but that lacks *Blochmannia*. In *L. niger*, we found that *vasa* protein (Vas), *nanos* mRNA (*nos*), and *oskar* mRNA (*osk*) localize in a single germplasm at the posterior pole, similar to other ants, wasps and flies (Fig. 1a–c). We found that these germline genes in *Camponotus floridanus*, a species in Camponotini that has a germplasm surrounded by *Blochmannia*, also localize in a single germplasm at the posterior pole in oocytes (Extended Data Fig. 1a–i). Surprisingly, we discovered that, as the oocyte transitions to a freshly laid egg, the mRNAs or proteins of nine germline genes localize in four subcellular locations that we name 'zones': zone 1 (the ancestral position of the germplasm at the posterior pole at 100% length of the egg), zone 2 (located at about 80% egg length), zone 3 (located at about 60% egg length) and zone 4 (at the anterior

[1]Department of Biology, McGill University, Montreal, Quebec, Canada. [2]Present address: Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Istanbul, Turkey. [3]These authors contributed equally: Ab. Matteen Rafiqi, Arjuna Rajakumar. ✉e-mail: ehab.abouheif@mcgill.ca

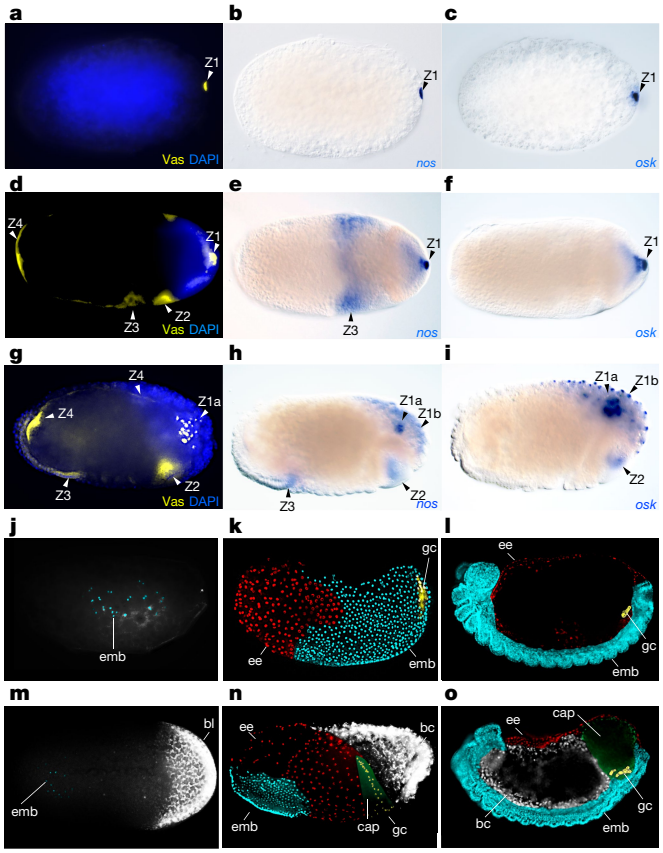


Fig. 1 | The evolution of four subcellular localization zones of germline genes that radically alter embryogenesis in *C. floridanus*. **a**–**c**, *L. niger* stage-1 freshly laid eggs, showing localization of Vas protein (**a**) in yellow, and *nos* mRNA (**b**) and *osk* mRNA (**c**) in blue. **d**–**f**, *C. floridanus* stage-1 freshly laid eggs, showing localization of Vas protein (**d**) in yellow, and *nos* mRNA (**e**) or *osk* mRNA (**f**) in blue. **g**–**i**, *C. floridanus* cellular blastoderm stage-6 embryos, showing expression of Vas protein (**g**) in yellow, and *nos* mRNA (**h**) and *osk* mRNA (**i**) in blue. **j**–**o**, Comparison of embryogenesis in *L. niger* (**j**–**l**) and *C. floridanus* (**m**–**o**). **j**, **m**, Freshly laid eggs (stage 1). **k**, **n**, Cellular blastoderm (stage 6). **l**, **o**, Segmentation (stage 12). False colouring highlights embryo (emb) in cyan, *Blochmannia* (bl) and bacteriocytes (bc) in white, germline (gc) in yellow, extraembryonic tissue (ee) in red and germline capsule (cap) in green. In **a**, **d** and **g**, blue indicates DAPI nuclear stain. Arrowheads indicate subcellular localization or expression zones of germline genes: zone (Z)1, zone 1a, zone 1b, zone 2, zone 3 and zone 4. Anterior is to the left; dorsal is to the top. In situ hybridization and immunohistochemistry experiments were repeated at least 4 (*L. niger*) or 8 times (*C. floridanus*) independently on $n \geq 5$ (*L. niger*) or $n \geq 30$ (*C. floridanus*) embryos per developmental stage.

pole at 0% egg length, extending along the dorsal side to the anterior boundary of *Blochmannia*) (Fig. 1d–f, Extended Data Fig. 1j–o). At a later stage, after the egg cellularizes and has initiated zygotic expression (the cellular blastoderm stage), the mRNAs or proteins of these nine genes persist in these four zones (Fig. 1g–i, Extended Data Fig. 1p–u). In both freshly laid and later-stage eggs, the localization and expression of mRNAs or proteins of germline genes is combinatorial—most are present in all four zones, but *nos* mRNA is only in two zones and *osk* mRNA is only in one zone (Fig. 1d–i, Extended Data Fig. 1j–u, Extended Data Table 1). Furthermore, the localization of these mRNAs or proteins is also dynamic: in later-stage eggs, the number of zones in which *nos* mRNA is present increases to three, and to two for *osk* mRNA—but the number of zones in which *smaug* (*smg*) mRNA is present decreases from four to three (Fig. 1d–i, Extended Data Fig. 1j–u, Extended Data Table 1). This combinatorial and dynamic localization shows that these four zones are not identical and suggests that they have distinct roles in integrating *Blochmannia* into *C. floridanus* during embryogenesis. Finally, because the freshly laid egg is a single host cell, the evolution of these four distinct zones is the result of changes in the subcellular localization of maternally inherited mRNAs and proteins.

We next asked how the evolution of the four zones of germline genes has affected embryogenesis in *C. floridanus*. We discovered that the eggs of *C. floridanus* are radically altered relative to those of other insects (Fig. 1j–o). Although insect embryos typically form at the posterior or throughout the entire egg[30,32], the embryo of *C. floridanus* forms in the anterior. At the posterior of the egg, *Blochmannia* become enveloped by specialized cells known as bacteriocytes and eventually migrate to the midgut to provide nutrition[11,14,16] (Fig. 1k, m–o, Extended Data Fig. 2a–c). Adjacent to bacteriocytes, the germ-cell precursors and a small population of *Blochmannia* are enveloped by a novel cell type we term the 'germline capsule', which—to our knowledge—has never previously been observed in insects (Fig. 1n, Extended Data Fig. 2d). The germline capsule then migrates posteriorly and attaches to the elongated embryo, where the germ cells and the small population of *Blochmannia* are transmitted to the next generation (Fig. 1o, Extended Data Fig. 2e, f). These results suggest that the four zones evolved to radically alter embryogenesis to integrate *Blochmannia* into *C. floridanus*.

We therefore investigated the role of each zone in this integration by tracking their fate in fixed embryos of known stages using Vas protein and *osk* mRNA (Extended Data Fig. 3). In freshly laid eggs, zone 1 initially appears as if it will form a posteriorly localized germplasm, as in other ants (Extended Data Fig. 3j, v). However, smaller germplasm foci begin budding off zone 1 and eventually give rise to two subzones: 'ancestral germplasm' at the posterior pole (which we term zone 1a) and germplasm foci at the centre of each bacteriocyte (zone 1b) (Extended Data Fig. 3a, b, j′, j″, m–p, v′, v″, w). At later stages, zone 1a and zone 1b migrate dorsally and are then no longer detectable (Extended Data Fig. 3c–h, p–u). By contrast, during the cellular blastoderm stage zone 2 is enveloped by the germline capsule and later migrates to connect to the embryonic posterior, where it gives rise to germ cells (Extended Data Fig. 3d–i, k, q–u, x). This shows that zone 2 is a novel germline of *C. floridanus*, and that the ancestral germplasm in zone 1 has lost its role in germline formation and acquired an alternative role within bacteriocytes. Next, zone 3 begins as a stripe in freshly laid eggs, and is later expressed throughout the germband, becoming enriched along the midline of the embryo; this suggests that zone 3 patterns the embryonic midline (Extended Data Fig. 3a–f). At this stage, zone 3 is also enriched at the posterior of the embryo, which suggests that it also specifies the embryonic posterior (Extended Data Fig. 3c–f). Finally, zone 4 is at the anterior pole in freshly laid eggs and then begins to extend dorsally, connecting to *Blochmannia* (Extended Data Fig. 3a, b). Later, zone 4 appears in the yolk membrane abutting the anteriormost cells of the embryo and extends all the way into the bacteriocytes (Extended Data Fig. 3c–e, l). Eventually, the yolk membrane forms the midgut that houses the bacteriocytes (Fig. 1o, Extended Data Fig. 2c). This suggests that zone 4 has a role in the migration of bacteriocytes to the midgut. Altogether, our data show that the four zones have distinct roles during the developmental integration of *Blochmannia* into *C. floridanus*—zone 1 and zone 4 have roles that are related to bacteriocytes; zone 2 is the functional germline; and zone 3 has a role in the embryonic midline and posterior. Zone 1 and zone 2 may have evolved to segregate *Blochmannia* into bacteriocytes for nutrition (zone 1) and into the germline capsule for vertical transmission (zone 2). Furthermore, zone 3 may have evolved to enhance the efficiency of this endosymbiosis by giving rise to an embryonic posterior in the anterior of the egg that is spatially separated from the *Blochmannia* populations in the posterior of the egg.

The Hox genes *abdA* and *Ubx* are rewired

In arthropods, the Hox genes *abdA* and *Ubx* function to morphologically differentiate the abdominal and thoracic segments after their

formation[33,34]. In hemipterans (from which the ancestor of *Blochmannia* colonized the Camponotini[10]), *abdA* and *Ubx* have an additional role in the development of bacteriocytes[35,36]. We therefore asked whether *abdA* and *Ubx* have a role in the integration of *Blochmannia* into *C. floridanus*. We discovered that—unlike in any other known insect— the mRNAs and proteins of *abdA* and *Ubx* in *C. floridanus* are localized in oocytes and freshly laid eggs, which shows that they are maternally inherited (Fig. 2a, d, Extended Data Fig. 4a–c). In freshly laid eggs, mRNAs and proteins of *abdA* localize in zone 1 and zone 3 and those of *Ubx* localize in all four zones (Fig. 2a, d, Extended Data Fig. 4c). At later stages, mRNAs and proteins of both genes are co-expressed with Vas protein in all four zones, and are also expressed in bacteriocytes (Fig. 2b, e, Extended Data Fig. 4d). Towards the end of embryogenesis, the conserved expression of *abdA* and *Ubx* appears in the third thoracic and abdominal segments (Fig. 2c, f, Extended Data Fig. 4e). Our results suggest that the mRNAs and proteins of *abdA* and *Ubx* interact with germline genes and have a role in all four zones.

To test this, we performed RNA interference (RNAi) on freshly laid eggs to knock down maternal and zygotic *abdA* and *Ubx* expression. We performed *abdA* RNAi at two concentrations, which produced a range of phenotypes in each zone as compared to a *YFP* RNAi control (Fig. 2g, h, j–n, Extended Data Fig. 4f–p). *abdA* RNAi at the lower concentration results in truncation of the embryonic posterior (zone 3) after the third abdominal segment (Fig. 2g, h). *abdA* RNAi at the higher concentration results in mild and severe phenotypes. Embryos with mild phenotypes develop into y-shaped embryos split along the ventral midline, and which—at later stages—truncate after the third abdominal segment (Fig. 2j–l, Extended Data Fig. 4f–k). Embryos with severe phenotypes develop into an embryonic stub or are undetectable (Extended Data Fig. 4l–p). Furthermore, *Ubx* RNAi results in the truncation of the embryonic posterior (zone 3) at the third thoracic segment (Fig. 2g, i). These results show that *abdA* and *Ubx* specify the embryonic posterior, and that *abdA* additionally functions in patterning the embryonic midline and forming the germband. Finally, *abdA* RNAi and *Ubx* RNAi also affect zone 1, zone 2 and zone 4: *Blochmannia* and bacteriocytes (zone 1 and zone 4) are eliminated (with *abdA* RNAi) or misplaced (with *Ubx* RNAi); the capsule (zone 2) develops external to the embryo (with *abdA* RNAi and with *Ubx* RNAi) or into an enlarged capsule (with *abdA* RNAi) (Fig. 2m–o, Extended Data Fig. 4o, p). Our RNAi data show that *abdA* and *Ubx* function in the four zones to integrate *Blochmannia* into *C. floridanus*.

We found that several germline genes were misexpressed after *abdA* RNAi, which suggests that *abdA* and *Ubx* are upstream of the germline genes (Extended Data Fig. 4f–n, p). To test this, we performed quantitative (q)PCR for nine germline genes on bacteriocytes (zone 1), the capsule (zone 2), and the germband and yolk sac together (zone 3 and zone 4) after dissecting them out of *YFP*-RNAi, *abdA*-RNAi, and *Ubx*-RNAi embryos (Extended Data Fig. 4q, r). We found that germline gene expression is downregulated in all three tissues after *abdA* RNAi and *Ubx* RNAi, and is significantly different from that in the *YFP*-RNAi control (Extended Data Fig. 4q, r). This shows that *abdA* and *Ubx* are rewired within the highly conserved segmentation hierarchy to regulate germline genes in the four zones.

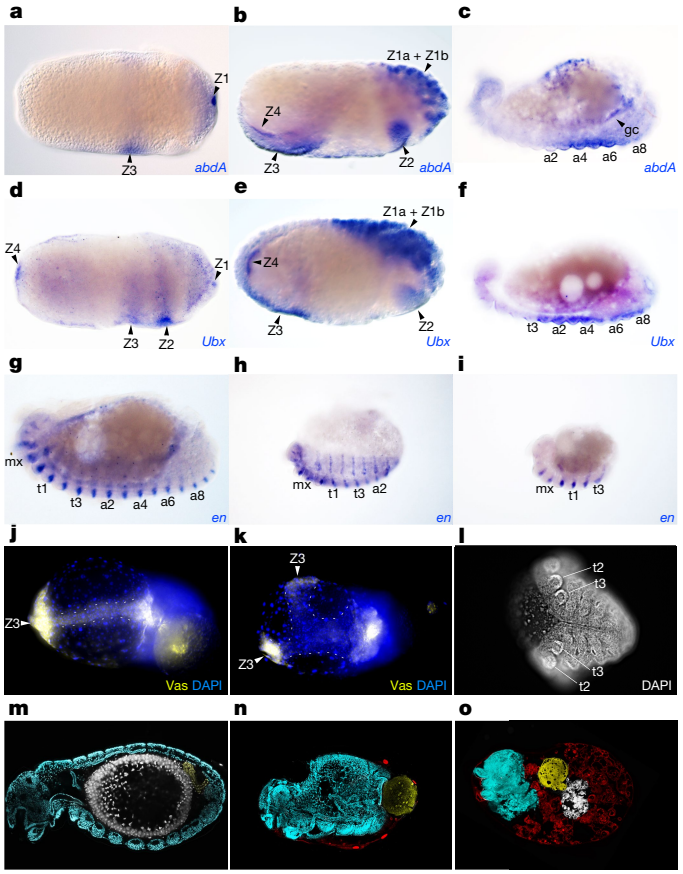


Fig. 2 | The Hox genes *abdA* and *Ubx* are rewired to regulate germline genes in *C. floridanus*. **a**–**f**, Wild-type *abdA* and *Ubx* mRNA staining in blue in freshly laid eggs (stage 1) (**a**, **d**), cellular blastoderm (stage 6) (**b**, **e**) and segmented (stage 12) (**c**, **f**) embryos. **g**–**i**, Stage-12 embryos with *engrailed* (*en*) staining in blue, showing the control *YFP* RNAi phenotype ($n = 70, 100\%$) (**g**), the low-concentration *abdA* RNAi phenotype ($n = 35$ out of 63, 56%) (**h**) and the high-concentration *Ubx* RNAi phenotype ($n = 113$ out of 122, 93%) (**i**). **j**, **k**, Stage-8 eggs with Vas protein in yellow, DAPI in blue and the embryo marked by dotted lines, showing the control *YFP* RNAi phenotype ($n = 45, 100\%$) (**j**) and the high-concentration *abdA* RNAi phenotype ($n = 21$ out of 61, 34%) (**k**). **l**, Stage-12 eggs, showing the high-concentration *abdA* RNAi phenotype with DAPI in white ($n = 22$ out of 31, 71%). **m**–**o**, Stage-17 eggs false-coloured to show embryo (cyan), serosa (red), *Blochmannia* and bacteriocytes (white), and germline capsule (yellow), showing the control *YFP* RNAi phenotype ($n = 70$, 100%) (**m**), the low-concentration *abdA* RNAi phenotype ($n = 35$ out of 63, 56%) (**n**) and the high-concentration *Ubx* RNAi phenotype ($n = 113$ out of 122, 93%) (**o**). Segments are marked as: maxillary (mx), thoracic segments (t)1–3 and abdominal segments (a)1–8. Zones and subzones are indicated with arrows. Anterior is to the left, dorsal is to the top; except for **j**–**l**, in which ventral is towards the reader. In situ hybridization and immunohistochemistry experiments (**a**–**f**) were repeated at least 8 times independently on $n \geq 30$ embryos per developmental stage.

Blochmannia regulate Hox and germline genes

Blochmannia makes up 97.2% of the total DNA content in freshly laid eggs (Extended Data Fig. 2g–k). We therefore tested whether *Blochmannia* influences the four zones. To do this, we treated *C. floridanus* colonies with rifampicin, an antibiotic that eliminates *Blochmannia*[11,14,15,19] (Fig. 3a, b, Extended Data Fig. 5a, c). We discovered that the formation of all four zones in freshly laid eggs is unaffected (Fig. 3c, Extended Data Fig. 6a, g, j, m, p). However, *abdA* mRNA and *tudor* protein (Tud) are lost from zone 1 and the ancestral germplasm becomes more tightly localized at the posterior pole, resembling the germplasm in other ants, wasps and flies[29–31] (Fig. 3a, f, Extended Data Fig. 6a, d, g, j, m, p). Therefore, in freshly laid eggs, all four zones are established by *C. floridanus*, while *Blochmannia* selectively regulates mRNAs and proteins to modify zone 1. At the cellular blastoderm stage, we observed a range of phenotypes that we categorized into two classes ('severe' and 'mild') that occur in equal proportion. Severe phenotypes are nonviable because they have no developing germband, show a range of morphological defects, and in all four zones, the mRNAs or proteins of *abdA*, *Ubx* and the germline genes are either absent or mislocalized (Fig. 3e, h, Extended Data Fig. 6c, f, i, l, o, r). By contrast, mild phenotypes are viable and show no morphological defects, and the mRNAs

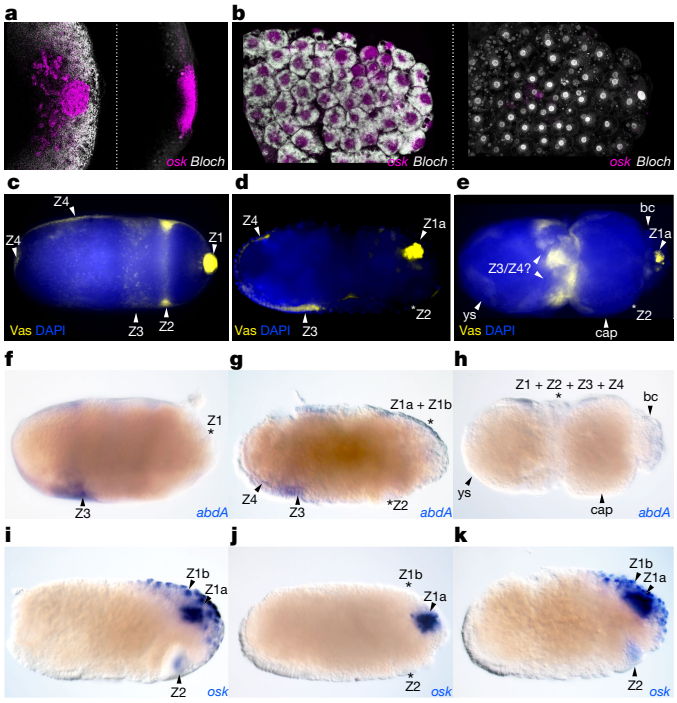


Fig. 3 | *Blochmannia* bacteria maintain and selectively regulate mRNA and proteins of maternal Hox and germline genes. a, Stage-1 freshly laid eggs stained with *osk* mRNA in magenta and DAPI, which marks *Blochmannia* (*Bloch*) in white, showing wild-type phenotype ($n \geq 30$) (left) and rifampicin-treated phenotype ($n \geq 15$) (right). **b**, Bacteriocytes from stage-6 embryos stained with *osk* mRNA in magenta and DAPI in white, showing wild-type phenotype ($n \geq 30$) (left) and rifampicin-treated phenotype ($n \geq 15$) (right). **c**–**e**, Embryos from rifampicin-treated colonies, showing Vas protein in yellow and DAPI in blue in a stage-1 freshly laid egg ($n = 29$) (**c**), a stage-6 embryo showing a mild phenotype ($n = 55$) (**d**) and a stage-6 embryo showing a severe phenotype ($n = 58$) (**e**). **f**–**h**, Embryos from rifampicin-treated colonies, showing *abdA* mRNA in blue in a stage-1 freshly laid egg ($n \geq 15$) (**f**), a stage-6 embryo showing a mild phenotype ($n \geq 15$) (**g**) and a stage-6 embryo showing a severe phenotype ($n = 6$) (**h**). **i**–**k**, Comparison of morphology and *osk* mRNA expression (blue) between stage-6 wild-type embryos (**i**), stage-6 embryos with a mild phenotype ($n = 39$) (**j**) and stage-6 embryos transplanted with *Blochmannia* ($n = 35$) (**k**), which were collected from the same *Blochmannia*-free rifampicin-treated colony as in **j**. One hundred per cent of the transplanted embryos develop into phenotypes similar to wild type ($n = 35$ of 35; compare **i** with **k**), and 0% ($n = 0$ of 35) develop into mild or severe phenotypes, one-tailed Fisher's exact test (degrees of freedom = 1, $P = 0.00002$). Asterisks indicate absence of localization or expression in zones or subzones. Arrowheads indicate zones or subzones. bc, bacteriocytes; cap, giant capsule; ys, yolk sac. Question marks indicate presumptive zones. Anterior is to the left, dorsal is to the top. In situ hybridization and immunohistochemistry experiments were repeated at least eight times independently.

or proteins of *abdA*, *Ubx* and the germline genes are selectively lost in each of the four zones: *abdA* and Tud in zone 1a; *abdA*, *Ubx*, *osk*, *nos* and Tud in zone 1b; *abdA*, *Ubx*, *osk*, *nos*, Vas and *aubergine* protein (Aub) in zone 2; and *staufen* mRNA (*stau*) in zone 3 and zone 4 (Fig. 3d, g, Extended Data Fig. 6b, e, h, k, n, q, Extended Data Table 1). At later stages, mild-phenotype embryos develop normally but often show defects in the gonads (Extended Data Fig. 5j, l, m, o–q). These results show that *Blochmannia* bacteria maintain mRNAs or proteins within each zone and selectively regulate them to make each zone functionally distinct.

To rule out the possibility that these changes are the unspecific effect of antibiotic treatment, we transplanted *Blochmannia* from wild-type eggs into *Blochmannia*-free eggs from a rifampicin-treated colony. One hundred per cent of the transplanted embryos developed into embryos similar to those of wild-type eggs, with *osk* mRNA restored to both zone 1b and zone 2; by contrast, un-transplanted control embryos from the same rifampicin-treated colony developed mild and severe phenotypes (Fig. 3i–k). Furthermore, we treated *C. floridanus* with ampicillin and a *Blochmannia*-free species (*L. niger*) with rifampicin. These embryos developed into embryos that are similar to those of the wild type in each case (Extended Data Fig. 5). Therefore, our results show that, after all four zones are established by *C. floridanus*, *Blochmannia* bacteria selectively regulate the mRNAs and proteins of *abdA*, *Ubx* and the germline genes to make the four zones functionally distinct and to maintain these zones. Because *abdA* and *Ubx* are upstream of the germline genes, *Blochmannia* selectively regulate germline genes through *abdA* and *Ubx*.

The origin and elaboration of integration

On the basis of our results, we predicted that the radical alterations we observed in *C. floridanus* embryos evolved in the most recent common ancestor of the Camponotini during the origin of the obligate endosymbiosis with *Blochmannia*. To test this prediction, we used RevBayes[37] to reconstruct the ancestral states of five developmental characters within the subfamily Formicinae (Fig. 4a, Extended Data Figs. 7, 8) to uncover the origin and elaboration of the developmental integration of *Blochmannia* into the Camponotini. In the ancestors of basally branching lineages, our reconstruction infers an embryo located in the posterior of the egg with a single germplasm (Fig. 4a (nodes 0–9 and 12–16), Extended Data Fig. 7 w–y, za, zb, zc, zd, ze). Notably, at node 10, node 11 and in *Brachymyrmex patagonicus*, the embryo shifted its location to the anterior but retained a single germplasm at the posterior of the egg (Fig. 4a, Extended Data Fig. 7t–v, z). Furthermore, the most recent common ancestor of the four closest sister tribes of the Camponotini evolved a novel subcellular localization zone for Vas and the maternal AbdA and Ubx proteins (Fig. 4a (nodes 12–16), Extended Data Figs. 7o–s, zf, zg, 8l–p). We infer that this zone is homologous to zone 3, because it is in a position similar to zone 3 in *C. floridanus* embryos and lacks *osk* mRNA (which exclusively marks zone 1 and zone 2 in *C. floridanus*) (Extended Data Fig. 7zh, zi, zj, zk). Finally, in addition to the Camponotini, different obligate endosymbionts evolved independently in the most recent common ancestor of the Formicini and Plagiolepidini[10,25,38] (Fig. 4a, Extended Data Fig. 7zf′, zg′). Therefore, the ability to shift the embryo to the anterior and the capacity to localize mRNAs and proteins to novel zones evolved before the three known obligate endosymbioses in ants at node 12 (Fig. 4a).

At the origin of the obligate endosymbiosis between *Blochmannia* and Camponotini, our reconstruction infers the evolution of three innovations: maternal AbdA and Ubx now localize to the ancestral germplasm (zone 1); zone 4 appears at the anterior pole and localizes Vas and maternal AbdA and Ubx; and the embryo shifts to the anterior of the egg, forming a novel embryonic posterior within zone 3 (Fig. 4a (node 17), Extended Data Figs. 7f–n, 8e–k). This integration was later elaborated within the derived genus *Camponotus* with two additional innovations: a germline in zone 2 that localizes Vas and maternal AbdA and Ubx and its surrounding germline capsule (Fig. 4a, Extended Data Figs. 7a–e, 8a–d). Our reconstruction uncovers the innovations that evolved before, during and after the origin of the obligate endosymbiosis between *Blochmannia* and Camponotini.

Discussion

Here we provide evidence for the following pathway for the origin and elaboration of developmental integration between *Blochmannia* and Camponotini (Fig. 4b). In step 1 (pre-existing capacity), a novel zone (zone 3) evolved to have a role in embryonic patterning, before the origin of this developmental integration. This led to a pre-existing capacity to localize mRNAs and proteins to novel subcellular locations,

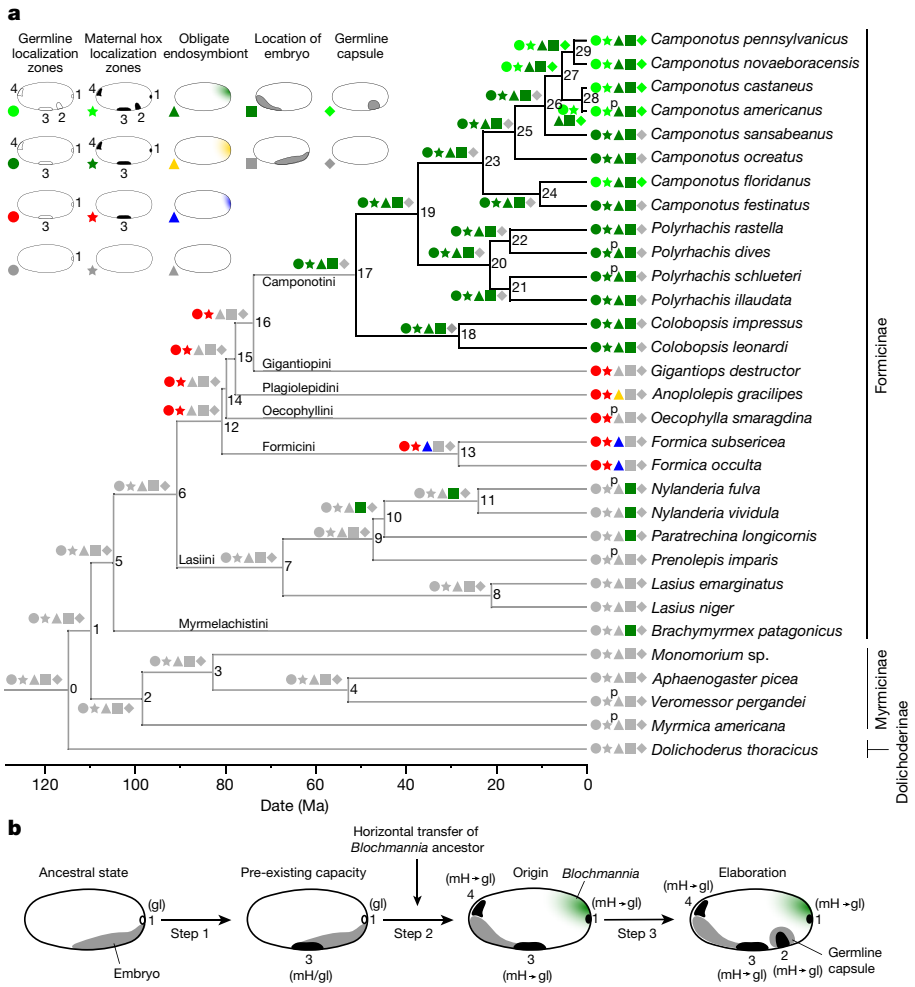


Fig. 4 | Origin and elaboration of developmental integration of *Blochmannia* into Camponotini. a, Phylogenetic tree of species within the ant subfamily Formicinae and outgroups; scale indicates millions of years ago (Ma). Nodes are numbered from 0 to 29. Species and subfamily names are indicated to the right of the tree, and names of tribes are indicated nearest to their node of origin. Black branches indicate lineages within the Camponotini with an obligate endosymbiosis with *Blochmannia*. Zones of mRNAs and proteins of germline genes are indicated by circles, and those of the maternal Hox genes *abdA* and *Ubx* are indicated by stars. Obligate endosymbionts in the posterior that are from different taxonomic lineages are indicated by triangles, the position of the embryo is indicated by a square and the presence of a germline capsule is indicated by a diamond. The different states of a developmental character are indicated by the different colours of the shaped icons. Character states predicted by RevBayes are marked with a 'p' superscript. Posterior probabilities for each node are listed in Extended Data Table 2. **b**, Proposed steps in the origin and elaboration of developmental integration between *Blochmannia* and the Camponotini. The number of localization zones (1 to 4) of mRNAs and proteins of germline genes (gl) and of the maternal Hox genes *abdA* and *Ubx* (mH) are indicated in black. Embryo and germline capsule are indicated in grey; *Blochmannia* is indicated in light green. Solidus (mH/gl) indicates that the maternal Hox genes are co-expressed with the germline genes, and arrows indicate that the maternal Hox genes are upstream of the germline genes.

which was subsequently co-opted to facilitate the integration of *Blochmannia* into the Camponotini. Furthermore, colocalization of *abdA*, *Ubx* and the germline genes in zone 3 facilitated the rewiring of *abdA* and *Ubx* to regulate the germline genes either before or at the origin of integration. In step 2 (origin), *Blochmannia* gained the ability to selectively regulate germline genes through *abdA* and *Ubx* in each zone, which led to the evolution of three functionally distinct zones: the ancestral germline (zone 1); the embryonic midline and posterior (zone 3), which allowed the shift of the embryo to the anterior; and zone 4, which guides bacteriocytes to the midgut. In step 3 (elaboration), some derived *Camponotus* species evolved a novel germline (zone 2) surrounded by a germline capsule, which freed the ancestral germline (zone 1) to have an alternative role within bacteriocytes. Therefore, the origin and elaboration of this major transition in individuality occurred through the stepwise addition of zones from 1 to 2 to 3 to 4 (Fig. 4b). This stepwise addition of zones evolved through tinkering with subcellular localization to produce distinct modules that divide labour within a single cell through the combinatorial localization of the same genes. Finally, the ecological mutualism between hemipteran bugs and the Camponotini[10,12,13] is thought to have facilitated the horizontal transfer of *Blochmannia* into the Camponotini, which suggests that ecological circumstances and pre-existing developmental capacities must converge to produce favourable conditions for major evolutionary transitions to obligate endosymbiosis. We therefore propose that other major transitions in individuality may originate and also elaborate through the rewiring of highly conserved gene regulatory networks, as well as by exploiting pre-existing molecular or developmental capacities and ecological interactions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2653-6.

1. Maynard-Smith, J. & Szathmary, E. *The Major Transitions in Evolution* (Oxford Univ. Press, 1997).
2. West, S. A., Fisher, R. M., Gardner, A. & Kiers, E. T. Major evolutionary transitions in individuality. *Proc. Natl Acad. Sci. USA* **112**, 10112–10119 (2015).
3. Sachs, J. L., Skophammer, R. G. & Regus, J. U. Evolutionary transitions in bacterial symbiosis. *Proc. Natl Acad. Sci. USA* **108** (Suppl 2), 10800–10807 (2011).
4. Boomsma, J. J. & Gawne, R. Superorganismality and caste differentiation as points of no return: how the major evolutionary transitions were lost in translation. *Biol. Rev. Camb. Philos. Soc.* **93**, 28–54 (2018).
5. Moran, N. A. Symbiosis as an adaptive process and source of phenotypic complexity. *Proc. Natl Acad. Sci. USA* **104** (Suppl 1), 8627–8633 (2007).
6. Ruiz-Trillo, I. & Nedelcu, A. M. *Evolutionary Transitions to Multicellular Life: Principles and Mechanisms* Vol. 2 (Springer, 2015).
7. Blochmann, F. Über das Vorkommen bakterienähnlicher Gebilde in den Geweben und Eiern verschiedener Insekten. *Zbl. Bakteriol.* **11**, 234–240 (1892).
8. Buchner, P. *Endosymbiosis of Animals with Plant Microorganisms* (Interscience, 1965).
9. Tanquary, M. C. *Biological and Embryological Studies on Formicidae*. PhD thesis, Univ. of Illinois (1912).
10. Wernegreen, J. J., Kauppinen, S. N., Brady, S. G. & Ward, P. S. One nutritional symbiosis begat another: phylogenetic evidence that the ant tribe Camponotini acquired *Blochmannia* by tending sap-feeding insects. *BMC Evol. Biol.* **9**, 292 (2009).
11. Zientz, E., Beyaert, I., Gross, R. & Feldhaar, H. Relevance of the endosymbiosis of *Blochmannia floridanus* and carpenter ants at different stages of the life cycle of the host. *Appl. Environ. Microbiol.* **72**, 6027–6033 (2006).
12. Aranda-Rickert, A., Fracchia, S., Yela, N. & Marazzi, B. Insights into a novel three-partner interaction between ants, coreids (Hemiptera: Coreidae) and extrafloral nectaries: implications for the study of protective mutualisms. *Arthropod-Plant Interact.* **11**, 525–536 (2017).
13. Clark, R. E., Farkas, T. E., Lichter-Marck, I., Johnson, E. R. & Singer, M. S. Multiple interaction types determine the impact of ant predation of caterpillars in a forest community. *Ecology* **97**, 3379–3388 (2016).
14. Feldhaar, H. et al. Nutritional upgrading for omnivorous carpenter ants by the endosymbiont *Blochmannia*. *BMC Biol.* **5**, 48 (2007).
15. de Souza, D. J., Bézier, A., Depoix, D., Drezen, J. M. & Lenoir, A. *Blochmannia* endosymbionts improve colony growth and immune defence in the ant *Camponotus fellah*. *BMC Microbiol.* **9**, 29 (2009).
16. Gil, R. et al. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA* **100**, 9388–9393 (2003).
17. Kupper, M., Stigloher, C., Feldhaar, H. & Gross, R. Distribution of the obligate endosymbiont *Blochmannia floridanus* and expression analysis of putative immune genes in ovaries of the carpenter ant *Camponotus floridanus*. *Arthropod Struct. Dev.* **45**, 475–487 (2016).
18. Ramalho, M. O., Vieira, A. S., Pereira, M. C., Moreau, C. S. & Bueno, O. C. Transovarian transmission of *Blochmannia* and *Wolbachia* endosymbionts in the neotropical weaver ant *Camponotus textor* (Hymenoptera, Formicidae). *Curr. Microbiol.* **75**, 866–873 (2018).
19. Sauer, C., Dudaczek, D., Hölldobler, B. & Gross, R. Tissue localization of the endosymbiotic bacterium "*Candidatus* Blochmannia floridanus" in adults and larvae of the carpenter ant *Camponotus floridanus*. *Appl. Environ. Microbiol.* **68**, 4187–4193 (2002).
20. Sauer, C., Stackebrandt, E., Gadau, J., Hölldobler, B. & Gross, R. Systematic relationships and cospeciation of bacterial endosymbionts and their carpenter ant host species: proposal of the new taxon *Candidatus* Blochmannia gen. nov. *Int. J. Syst. Evol. Microbiol.* **50**, 1877–1886 (2000).
21. Wolschin, F., Hölldobler, B., Gross, R. & Zientz, E. Replication of the endosymbiotic bacterium *Blochmannia floridanus* is correlated with the developmental and reproductive stages of its ant host. *Appl. Environ. Microbiol.* **70**, 4096–4102 (2004).
22. Ward, P. S., Blaimer, B. B. & Fisher, B. L. A revised phylogenetic classification of the ant subfamily Formicinae (Hymenoptera: Formicidae), with resurrection of the genera *Colobopsis* and *Dinomyrmex*. *Zootaxa* **4072**, 343–357 (2016).
23. Sinotte, V. M., Freedman, S. N., Ugelvig, L. V. & Seid, M. A. *Camponotus floridanus* ants incur a trade-off between phenotypic development and pathogen susceptibility from their mutualistic endosymbiont *Blochmannia*. *Insects* **9**, 58 (2018).
24. Stoll, S., Feldhaar, H., Fraunholz, M. J. & Gross, R. Bacteriocyte dynamics during development of a holometabolous insect, the carpenter ant *Camponotus floridanus*. *BMC Microbiol.* **10**, 308 (2010).
25. Sameshima, S., Hasegawa, E., Kitade, O., Minaka, N. & Matsumoto, T. Phylogenetic comparison of endosymbionts with their host ants based on molecular evidence. *Zool. Sci.* **16**, 993–1000 (1999).
26. Degnan, P. H., Lazarus, A. B., Brock, C. D. & Wernegreen, J. J. Host–symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of *Camponotus* species and their endosymbionts, *Candidatus* Blochmannia. *Syst. Biol.* **53**, 95–110 (2004).
27. Extavour, C. G. & Akam, M. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* **130**, 5869–5884 (2003).
28. Lehmann, R. & Nüsslein-Volhard, C. The maternal gene *nanos* has a central role in posterior pattern formation of the *Drosophila* embryo. *Development* **112**, 679–691 (1991).
29. Lehmann, R. Germ plasm biogenesis—an oskar-centric perspective. *Curr. Top. Dev. Biol.* **116**, 679–707 (2016).
30. Khila, A. & Abouheif, E. Reproductive constraint is a developmental mechanism that maintains social harmony in advanced ant societies. *Proc. Natl Acad. Sci. USA* **105**, 17884–17889 (2008).
31. Lynch, J. A. et al. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.* **7**, e1002029 (2011).
32. Lynch, J. A. & Roth, S. The evolution of dorsal–ventral patterning mechanisms in insects. *Genes Dev.* **25**, 107–118 (2011).
33. Akam, M. Hox genes, homeosis and the evolution of segment identity: no need for hopeless monsters. *Int. J. Dev. Biol.* **42**, 445–451 (1998).
34. Hughes, C. L. & Kaufman, T. C. *Hox* genes and the evolution of the arthropod body plan. *Evol. Dev.* **4**, 459–499 (2002).
35. Braendle, C. et al. Developmental origin and evolution of bacteriocytes in the aphid–*Buchnera* symbiosis. *PLoS Biol.* **1**, e21 (2003).
36. Matsuura, Y., Kikuchi, Y., Miura, T. & Fukatsu, T. *Ultrabithorax* is essential for bacteriocyte development. *Proc. Natl Acad. Sci. USA* **112**, 9376–9381 (2015).
37. Höhna, S. et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).
38. Wernegreen, J. J., Degnan, P. H., Lazarus, A. B., Palacios, C. & Bordenstein, S. R. Genome evolution in an insect cell: distinct features of an ant–bacterial partnership. *Biol. Bull.* **204**, 221–231 (2003).

Methods

No statistical methods were used to predetermine sample size. The experiments that were randomized are indicated below and investigators were not blinded to allocation during experiments and outcome assessment, with the exception of the qPCR experiment, in which the technician at the IRIC-Genomics Platform was blinded to allocation during experiments and outcome assessment.

Ant culturing and collection

Colonies were maintained in plastic boxes with glass test tubes filled with water constrained by cotton wool, and were fed a combination of mealworms, crickets, fruit flies and Bhatkar–Whitcomb diet[39]. All colonies were maintained at 25 °C, 70% relative humidity and a 12-h day:night cycle.

Colonies were collected from the following locations: *Aphaenogaster picea*, *Camponotus pennsylvanicus*, *Formica subscericea* and *Lasius niger* were collected at McGill Gault Nature Reserve (Quebec, Canada) (45° 32′ 12.4″ N, 73° 09′ 10.1″ W ± 1 km). *Camponotus novaeboracensis* ants were collected at Winnipeg (Manitoba, Canada) (49° 51′ 12.6″ N, 97° 08′ 14.0″ W ± 1 km). *Camponotus floridanus*, *Camponotus castaneus* and *Monomorium* sp. were collected at Gainesville (Florida, USA) (29° 42′ 05.7″ N, 82° 20′ 43.5″ W ± 1 km). *Colobopsis impressus* ants were collected at Gainesville (Florida, USA) (29° 41′ 07.4″ N, 82° 13′ 38.5″ W ± 1 km). *Camponotus ocreatus*, *Camponotus sansabeanus*, *Formica occulta* and *Veromessor pergandei* were collected at Miami (Arizona, USA) (33° 24′ 28.1″ N, 111° 00′ 14.5″ W ± 1 km). *Camponotus festinatus*, *Camponotus americanus*, *Camponotus sansabeanus*, *Brachymyrmex patagonicus*, *Nylanderia fulva* and *Nylanderia vividula* were collected at University of Texas at Austin, Brackenridge Field Laboratory (Texas, USA) (30° 17′ 2.40″ N, 97° 46′ 40.80″ W ± 1 km). *Myrmica americana* and *Prenolepis imparis* were collected at Medford (New York, USA) (40° 48′ 6.8566″ N, 73° 0′ 16.7756″ W ± 1 km). *Gigantiops destructor* ants were collected at ACTS Research Station (Maynas, Peru) (3° 14′ 60.00″ S, 72° 54′ 36.00″ W ± 1 km). *Anoplolepis gracillipes*, *Dolichoderus thoracicus*, *Oecophylla smaragdina*, *Paratrechina longicornis*, *Colobopsis leonardi* and *Polyrhachis rastellata* were collected at Mae Tang (Chiang Mai, Thailand) (location data not available). *Polyrhachis schlueteri* were collected at Bela Bela (Limpopo, South Africa) (24° 47′ 32.0″ S, 28° 17′ 30.6″ E ± 1 km). *Polyrhachis illaudata* and *Polyrhachis dives* were collected at Hong Kong region Guangdong (China) (location data not available). *Lasius emarginatus* ants were collected at Palmanova (Udine, Italy) (45° 54′ 31.5″ N, 13° 18′ 45.2″ E ± 1 km).

Aphaenogaster picea, *Camponotus pennsylvanicus*, *Formica subscericea*, *Lasius niger*, *Camponotus castaneus*, *Camponotus floridanus*, *Colobopsis impressus*, *Monomorium* sp., *Camponotus ocreatus*, *Camponotus sansabeanus*, *Formica occulta*, *Veromessor pergandei*, *Camponotus festinatus*, *Camponotus americanus*, *Camponotus sansabeanus*, *Brachymyrmex patagonicus*, *Nylanderia vividula*, *Myrmica americana* and *Prenolepis imparis* were collected by the laboratory of E.A. *Camponotus novaeboracensis* was collected by J. Rand, *Nylanderia fulva* was collected by E. Lebrun and *Gigantiops destructor* was collected by J. Gibson (laboratory of A. Suarez). *Anoplolepis gracillipes*, *Dolichoderus thoracicus*, *Oecophylla smaragdina*, *Paratrechina longicornis*, *Colobopsis leonardi* and *Polyrhachis rastellata* were purchased from Ants of Asia (P. Williams), and *Polyrhachis schlueteri*, *Polyrhachis illaudata*, *Polyrhachis dives* and *Lasius emarginatus* were purchased from Antstore (M. Sebesta).

Ovary dissections

This protocol was modified from a previous publication[40]. Ovaries were dissected in 0.1% PBSTween (1.86 mM NaH_2PO_4, 8.41 mM Na_2HPO_4, 1.75 M NaCl, 0.1% Tween20, pH 7.4) and kept on ice until fixation. First, the ovaries were removed from the oviduct. Ovaries were then separated into individual ovarioles, and the peritoneal sheath was then removed with fine forceps. Ovarioles were fixed in a solution of 5% formaldehyde (135 µl), 10% DMSO (100 µl) in 0.1% PBSTween (765 µl) for 25 min at room temperature, then washed with 0.1% PBSTween and gradually transferred to a solution of 100% methanol for storage.

Embryo collection and fixation

This protocol was modified from previous publications[40–42]. Embryos were treated with 4% hypochlorite solution (bleach) for 2 min. Embryos used for immunohistochemistry were then fixed using a 'slow formaldehyde fixing method' using PEMS (100 mM PIPES, 2 mM $MgSO_4$, 1 mM EGTA, pH 6.9) and were treated with proteinase K (New England Biolabs) in PBS at a final concentration of 0.08 U/ml. Embryos used for in situ hybridization were heat-fixed using a boiling hot solution of PBS-Triton (1.86 mM NaH_2PO_4, 8.41 mM Na_2HPO_4, 1.75 M NaCl, 0.03% Triton-X-100, pH 7.4).

Embryo staging

Timed egg depositions were collected in smaller setups and allowed to develop at 25 °C, 70% relative humidity along with a few workers, and fixed at two-hour windows. The embryos were DAPI-stained and observed under differential interference contrast (DIC) and wide-field fluorescence for staging. As far as possible, the staging scheme landmarks used correspond to Bownes' staging scheme for *Drosophila*[43].

Whole-genome shotgun sequencing

Whole DNA was isolated from 0–6-h-old embryos using Qiagen Genomic-tip 20/G kit. Shotgun sequencing was performed at Genome Quebec using Illumina HiSeq platform. Sequences were curated and BLAST searches performed using Geneious software[44].

Gene cloning and molecular biology

Gene sequences were obtained from NCBI GenBank database using genome BLAST against the assembled *C. floridanus* genome[45]. The accession numbers of genes used in this study are: *abdA* XM_020027891.2; *nos* XM_011266396; *osk* XM_011254572.2; *smg* XM_011254071.3; *stau* XM_011254361.3; *Ubx* XM_011259757.1; *en* XM_011252307.3. It was necessary to use a better-annotated *Ubx* cDNA sequence, which was submitted to GenBank under accession number MH801205. *Camponotus floridanus* and *L. niger* RNA was isolated using TRIzol (Invitrogen) from a pool of embryos and larvae of different developmental stages. RNA was then reverse-transcribed to synthesize a cDNA library. Specific primers were designed to amplify the gene fragments from cDNA libraries prepared from embryos and cloned in pGemT-easy vector (Promega) using standard procedures, and subsequently sequenced using Sanger sequencing at the Genome Quebec Innovation Centre. The primers used were: *osk* forward 5′-CGGAGAGCCTATTCCTTATC-3′, and reverse 5′-GCCAGAGATCTGATCCAATTA-3′, *nos* forward 5′-TCCCAGT TTGGACGAAGAATAAAG- 3′, and reverse 5′-GTTTTCCCGCAGAG TTTCTCAGTA- 3′, *stau* forward 5′-GCGAATTCACGGGTAGAGGT- 3′, and reverse 5′-GAAACACCAGCCGCATTCTG- 3′, *abdA* forward 5′-GTCTTC CTAAGAGCGACGAGC-3′, and reverse 5′-GTGGGTACCTTACTGACTGCC-3′, *Ubx* forward 5′-GCTTCTACGGAAGCCACCATC-3′, and reverse 5′-TGCTTC TCCTGCTCGTTTAGC-3′, *smg* forward 5′-TCACTTTTGCGTCGTCTACCT-3′, and reverse 5′-AGAGAGAGCCAGTTTGTGCC- 3′, *en* forward 5′-CGACACGAGCGAGGTATTGA- 3′, and reverse 5′-GAGGCCGATCGA TTTGACGA- 3′.

Identifying orthologues and paralogues

Amino acid sequence alignments were done using ClustalW in Geneious platform confirming the orthology of *vasa* (*vas*), *oskar* (*osk*), *nanos* (*nos*), *tudor* (*tud*), *germ cell-less* (*gcl*), *staufen* (*stau*), *caudal* (*cad*), *smaug* (*smg*), *wunen-2* (*wun2*), *aubergine* (*aub*), *heat shock protein 90* (*hsp90*), *argonaute 3* (*ago3*), *abdominal A* (*abdA*), *Ultrabithorax* (*Ubx*) and *engrailed* (*en*). The alignments are presented in Supplementary Figures 1–15. To search for any lineage-specific paralogues of the

germline genes that we studied, a blastn search was performed on the latest *C. floridanus* genome assembly accessed from: https://www.ncbi.nlm.nih.gov/assembly/1752781 using a maximum *E* value of 0.05; scoring of 2, −3; and gap cost of 5, 2. Only hits above an *e*-value cut off (e^{-20}) were considered and highlighted in bold. If the query subject had more than one hit but aligned to the same contig number, then it was concluded that no paralogues for the gene in question exist. However, if the query subject had more than one hit but was aligned to multiple contig numbers, then it was concluded that paralogues for the gene in question do exist. The hit tables are presented as Supplementary Table 3.

Immunohistochemistry and in situ hybridization

The following are primary antibodies we used in this study, the concentrations at which we used them at, and their source: mouse anti-HSP90 (1:100) antibody (BD bioscience 610418) and mouse anti-UbdA (1:4) antibody (FP6.87, DSHB), rabbit anti-Vasa (1:100) antibody (gift from P. Lasko), rabbit anti-Tudor (1:100) antibody (gift from P. Lasko), rabbit anti-Germ cell-less (1:300) antibody (gift from P. Lasko), rabbit anti-Aubergine (1:50) antibody (gift from P. Lasko), and rabbit anti-Oskar (1:100) antibody (gift from P. Lasko). Fluorescent secondary donkey anti-rabbit and anti-mouse polyclonal Alexa Fluor-488 (AbCam) antibodies were used at 1:500 dilution to detect the primary antibody, according a previous publication[30]. In situ hybridization was done according to previous publications[40,46], modified for in situ robot InsituPro VSi (Intavis) with the following modifications; the duration of wash steps was maintained according to the cited protocol but the buffer was exchanged every 5 min to increase agitation. Alkaline phosphatase secondary antibody anti-DIG-AP (Roche) was used to detect DIG-labelled probes and streptavidin-AP (Roche) reagent was used to detect biotin-labelled probes. Templates for probes were prepared using PCR with T7 and SP6 primers on the plasmids containing cloned gene fragments. Probe synthesis was done using SP6 or T7 RNA polymerase (Roche) according to the suppliers' directions. Probes were purified using phenol–chloroform and isopropanol precipitation method according to a previous publication[47], and used at 3 ng/µl final concentration. The probes consisted of 538 bp of *abdA* (bases 28–566), 424 bp of *nos* (bases 51–475), 848 bp of *osk* (bases 96–944), 987 bp of *stau* (bases 1121–2108) and 874 bp of *Ubx* (bases 32–906), 1,037 bp of *en* (bases 201–1237) and 949 bp of *smg* (bases 268–1197), in which base numbering starts at the start codon.

Microinjections and RNAi for phenotypic analysis

Embryos were collected as timed depositions from queens isolated with at least a dozen minor workers and at least six larvae and pupae. To eliminate any colony or day-of-injection related effects, embryos from multiple queens were collected, randomized between treatment and control and injected on the same day. Embryos were lined up alongside a fine glass capillary on a Petri dish lid lined with a thin layer of 2% agar in water and supplemented with 10 µl of 10 µg/ml ampicillin, modified from a previous publication[48]. Injection needles were prepared using a micropipette capillary puller. Microinjections were done using FemtoJet Express and InjectmanNI2 (Eppendorf) setup on a Zeiss Axiovert zoom inverted microscope using the following settings: control pressure 2 psi, injection pressure 18 ps, and injection time 0.1 s. The needle tip was broken open by gently pushing it against a glass coverslip immersed in halocarbon oil. The injection volume was adjusted after the needle was broken. Injection volumes were between 0.5 and 1 nl. Embryos were incubated at 25 °C 70% relative humidity chamber. Embryos were transferred every 24 to 48 h on to fresh 50-mm Petri-dishes containing 2% agar in water topped with a Whatmann filter paper and supplemented with 10 µl of 10 mg/ml ampicillin. DNA templates for double-stranded (ds) RNA were prepared using PCR with M13 forward universal primer and M13 reverse universal primer containing a T7 promoter overhang on plasmids containing cloned gene fragments as templates. The templates were used to generate dsRNA using T7 RNA polymerase (Roche) according to manufacturer's instructions. For controls, dsRNA was generated using the same method from a plasmid containing cloned 720 bp of the YFP coding sequence.

Quantitative PCR

Microinjections and RNAi. Embryos were collected as timed depositions from queens isolated with at least a dozen minor workers and at least six larvae and pupae. To eliminate any colony or day-of-injection related effects, embryos from multiple queens were collected, randomized between treatment and control and injected on the same day. Embryos were lined up alongside a fine glass capillary on a Petri dish lid lined with a thin layer of 2% agar in water and supplemented with 10 µl of 10 µg/ml ampicillin, modified from a previous publication[48]. Injection needles were prepared using a micropipette capillary puller. Microinjections were done using FemtoJet Express and InjectmanNI2 (Eppendorf) setup on a Zeiss Axiovert zoom inverted microscope using the following settings: control pressure 2 psi, injection pressure 18 psi and injection time 0.1 s. The needle tip was broken open by gently pushing it against a glass coverslip immersed in halocarbon oil. The injection volume was adjusted after the needle was broken. Injection volumes were between 0.5 and 1 nl. Embryos were incubated at 25 °C 70% relative humidity chamber. Embryos were transferred every 24 h on to fresh 50-mm Petri dishes containing 2% agar in water topped with a Whatmann filter paper and supplemented with 10 µl of 10 mg/ml ampicillin. DNA templates for dsRNA were prepared using PCR with M13 forward universal primer and M13 reverse universal primer containing a T7 promoter overhang on plasmids containing cloned gene fragments as templates. The templates were used to generate dsRNA using T7 RNA polymerase (Roche) according to manufacturer's instructions. For controls, dsRNA was generated using the same method from a plasmid containing cloned 720 bp of the YFP coding sequence.

Sample preparation for qPCR. Embryos were collected at stage 8, 5 days after injection and heat-fixed by immersing in a boiling hot solution of PBS-Triton (1.86 mM NaH_2PO_4, 8.41 mM Na_2HPO_4, 1.75 M NaCl, 0.03% Triton-X-100, pH 7.4) for 1 min followed by rinses with ice-cold PBS. Individual germline capsules, bacteriocytes and yolk sacs with intact germbands curled around them were separated using sharpened tungsten needles, and extraembryonic serosa tissue was discarded. For each gene, 40 individual samples were divided into 4 technical replicates of 10, and the 10 samples within each technical replicate were pooled and immediately placed in 200 µl Trizol reagent. RNA was prepared on the same day using standard Trizol method. First-strand cDNA synthesis was done using Superscript-II reverse transcriptase (ThermoFisher). Instead of universal oligo-dT primers an equimolar pool of the following 21 gene specific primers (including those of 8 endogenous control genes) was used to account for low yields in small tissue preparations: *vasa* 5′-CGATATCTGGTAGAAAGCCC-3′, *osk* 5′-GCCAGAGATCTGATCCAATTA-3′, *nos* 5′-GTTTTCCCGCAGAGTTTCTCAGTA- 3′, *tud* 5′-AGCGCCGGTTCTATCATGTC- 3′, *gcl* 5′-CCATCTCCAAGTATGTTCACC- 3′, *stau* 5′-GAAACACCAGCCGCATTCTG- 3′, *smg* 5′-AGAGAGAGCCAGTTTGTGCC- 3′, *ago3* 5′-TACACCCGTTATGCTTTTGA- 3′, *cad* 5′-AGAGGCGCCGATAGAGATGAA- 3′, *arm* 5′-TCTCGGTGCCTGTGATTCTG- 3′, *abdA* 5′-TCCAGGCCGCTTACGTGATG- 3′, *Ubx* 5′-TGCTTCTCCTGCTCGTTTAGC- 3′, *wun2* 5′-TCGTAATCGGTAGGTCGATGC- 3′, *act5c* 5′-GAACGGTGTTGGCGTACAGA- 3′, *tub* 5′-CGACGGAGAGTTGTTCGTGA- 3′, *argk* 5′-CCTGTCCAAGATCACCACCC- 3′, *ef1* 5′-AGTGGTCAATCCAGCAGGTG- 3′, *efl-like* 5′-GCAGCTGGTATTCCCGTTTG- 3′, *hisH-3* 5′-CCCTGAAAAGGGCCGATTGT- 3′, *rp60S* 5′-AACGTGCACTGGCATTTGTC- 3′, and *gadph* 5′-ATTCGCCATACGACGAGACC- 3′.

Quantitative PCR. Quantitative PCR was performed at IRIC-Genomics Platform using qPCR Taqman method[49] with the following primers: *vasa* forward 5′-CACAAGTACTTATTGTATCACCCACA-3′and

reverse 5′-GAAAATTTCTTGGCCTGTTGA-3′, *osk* forward 5′-AATCTCG TCGGAGAGCCTAT-3′and reverse 5′-AAATGCACGGAGACTCGAAA-3′, *nos* forward 5′-CCTTACCAACAGAATGCGTCT-3′and reverse 5′-TCCT TTAGCAGATGTTTTCGATAG-3′, *tud* forward 5′-ATTGTGGGTAC GAATATGTTATCG-3′and reverse 5′-ATGACAATGGTGTTAACATAAAGGAT -3′, *gcl* forward 5′-AAAACGATGGTTGGAAGTCAA-3′and reverse 5′-TGCCATTAAATCTGGTGCAA-3′, *stau* forward 5′-AACCCGCG AAACCATCTAT-3′and reverse 5′-CGTCACTTTTCTGGGTTTCG-3′, *ago-3*, forward 5′-TGGCATAGATGTCTATCATGCTG-3′and reverse 5′-GCAAC AAATCCTGCAACACTC-3′, *cad* forward 5′-ATGTCAATGCAGGCAGCAC-3′and reverse 5′-ACGTGGACGGAGATGTCG-3′, *wun2* forward 5′-TCTTGGC ACAATCGTAGCTTT-3′and reverse 5′-TCCGTGGAAGAATGCCTCT-3′, *tub* forward 5′-CACAGGCACGTATCGACAAC-3′and reverse 5′-GCCACGCG CATAATTGTT-3′, *act5c* forward 5′-CGTCATCAGGGTGTCATGG-3′ and reverse 5′-CAAGATACCTCTCTTCGATTGAGC-3′, *rp60S* forward 5′-GCGTTTCAAGGGCCAATAC-3′and reverse 5′-GCAGCATGTGA CGTGTTTTC-3′, *argk* forward 5′-TGGTAGACGCAGCGGTTT-3′and reverse 5′-AACGACTTGCTGTCGGATTC-3′, *efl-like* forward 5′-ACGTTATT GTCGAGGCCAAG-3′and reverse 5′-GGCAGGACGTATCTGCGTA-3′, *ef1* forward 5′-GCTGCAGTCGCATTTGTTC-3′and reverse 5′-ATCTTGGAAGAT GGCTCCAG-3′, *gapdh* forward 5′-GCGGTGCCAAGAAGGTTAT-3′and reverse 5′-CCAAGTTTACACCGACAACG-3′, *hisH-3*, forward 5′-CTACTAAAGCGG CGAGGAAG-3′and reverse 5′-CCAGGCCTATAACGATGAGG-3′.

The endogenous control genes used (the last 8 primer pairs above) were: *Actin5c*, *60S ribosomal protein*, *arginine kinase*, *efl-like*, *elongation factor1*, *gadph*, *histone H3* and *tubulin*. The most stable endogenous controls were established through the use of the algorithms integrated in the RefFinder package[50] that integrates four different protocols; GeNorm, BestKeeper, NormFinder and the comparative ΔC_T method[51–54] (Supplementary Fig. 16). Four of the endogenous controls (*gapdh*, *hisH3*, *rp60S* and *argk*) were deemed most stable and the geometric mean of these was used for calculating ΔC_T values for each target gene within each biological sample and replicate according to previously published recommendations[51] (Supplementary Fig. 16). Relative quantifications for *abdA* RNAi, *Ubx* RNAi and control *YFP* RNAi were calculated by the formula: relative quantification = $2^{-\Delta\Delta C_T}$, in which $\Delta\Delta C_T$ is the difference between ΔC_T in each RNAi sample and the average of ΔC_T values in all *YFP* RNAi replicates of that treatment group. $\Delta\Delta C_T$ values for each of the individual data points of the control *YFP* RNAi were also calculated using the average of all *YFP* RNAi from that particular biological replicate (black bars in Extended Data Fig. 4q, r). This method allows for consistency because the statistical analyses are performed on the same relative quantification values that are used to plot the bar graphs.

Antibiotic treatment

Two mature colonies were treated with rifampicin to test the effects of *Blochmannia* on embryonic development of *C. floridanus*. Rifampicin powder (Sigma; R883) was dissolved in water at a stock concentration of 2 mg/ml and then diluted 1:1 (final concentration 1 mg/ml rifampicin) in a 50% honey–water (Kirkland Signature) solution. Colonies were given fresh rifampicin–honey–water three times a week for two months. After two months, embryos were collected, fixed and stained with DAPI to confirm elimination of *Blochmannia*. Once elimination of *Blochmannia* was confirmed, embryos were collected and fixed for subsequent gene expression analysis. To rule out the possibility that the changes in phenotypes and gene expression or localization observed after rifampicin treatment are the unspecific effect of antibiotics were performed two controls: (1) a *C. floridanus* colony was treated with ampicillin, which does not eliminate *Blochmannia* from the colonies. Ampicillin powder (Fisher scientific; BP1760-25) was dissolved in water at a stock concentration of 400 mg/ml and then diluted 1:1 (final concentration 200 mg/ml ampicillin) in 50% honey–water solution. Colonies were treated in exactly the same manner as that for rifampicin. (2) An *L. niger* colony—a species that is in the same subfamily as *C. floridanus* but lacks *Blochmannia*—was treated with rifampicin in the same manner as *C. floridanus*. *Lasius niger* colonies were also treated with the same rifampicin regimen as *C. floridanus* and embryos were collected and fixed for subsequent gene expression analysis after at least two months.

Phylogenetic sampling, developmental characters and ancestral state reconstruction

Phylogenetic sampling. Thirty-one ant species were sampled in total: 26 from the subfamily Formicinae and 5 from 2 sister subfamilies of the Formicinae, the Myrmicinae (4 species) and the Dolichoderinae (1 species). Within the Formicinae, 14 in-group species within the Camponotini that evolved the obligate endosymbiosis with *Blochmannia* were sampled, and 12 out-group species were sampled that lack *Blochmannia*. Phylogenetic relationships and branch length information for these 31 species were obtained from previous molecular phylogenetic studies[26,55–57].

Developmental characters. The following five developmental characters were characterized for each species: (1) character 1 is defined as the presence of specific localization zones of mRNAs and proteins of the germline genes based on Vas protein. Character 1 has four states: an embryo with the presence of 1, 2, 3 or 4 localization zones of mRNAs and proteins of germline genes as illustrated in Fig. 4a; (2) character 2 is defined as the presence of specific localization zones of mRNAs and proteins of the maternal Hox genes *Ubx* and *abdA* based on the UbdA antibody that recognizes both Ubx and AbdA protein (with the exception of 1 species, *C. impressus*, which is based on *abdA* mRNA). Character 2 has four states: an embryo with the presence of 0, 1, 3 or 4 localization zones of mRNAs and proteins of maternal Hox genes *Ubx* and *abdA*, as illustrated in Fig. 4a; (3) character 3 is defined as the presence and type of obligate endosymbionts at the posterior of the egg on the basis of our own data and previous studies[8,9,58,59]. Previous phylogenetic evidence[10,25,38] showed that the three types of obligate endosymbionts within the Formicinae—the Camponotini obligate endosymbiont (*Blochmannia*), the Formica obligate endosymbiont and the Plagiolepidini obligate endosymbiont—were acquired independently and evolved convergently. Therefore, character 3 has four different states: an obligate endosymbiont at the posterior is absent; the Camponotini obligate endosymbiont (*Blochmannia*) is present at the posterior of the egg; the Formicini obligate endosymbiont is present at the posterior of the egg; or the Plagiolepidini obligate endosymbiont is present at the posterior of the egg, as illustrated in Fig. 4a; (4) character 4 is defined as the location of the embryo within the egg. Character 4 has two states: either the embryo is located in the posterior of the egg or the embryo is located in the anterior of the egg, as illustrated in Fig. 4a; (5) character 5 is defined as the germline capsule. Character 5 has two states: either the germline capsule is present or the germline capsule is absent, as illustrated in Fig. 4a.

Ancestral state reconstruction. RevBayes (v.1.7.10)[37] was used to reconstruct ancestral character states for the 5 developmental characters across 31 ant species sampled. RevBayes[37] uses Bayesian Markov chain Monte Carlo (MCMC) methods to estimate model parameters[60]. Ancestral states were estimated using two evolutionary models for discrete characters: the 'equal-transition rates' and 'unequal-transition rates' models. The equal-transition rates model assumes characters are equally likely to change from any one state to any other state, whereas the unequal-transition rates model assumes that the transition from any one state to any other state is unequal and can occur according to different rate parameters[37]. Both models were applied on each of the five developmental characters, and each model was run independently twice for 1,000,000 MCMC generations sampling every 500 generations. After completion of the MCMC analysis, the first 25% of the trees were discarded as a burn-in. Convergence between chains, likelihood scores and estimate sample size values were evaluated using Tracer

(version 1.7)[61]. The estimate sample size value for each parameter sampled from the MCMC analysis was always recorded as >1,000, indicating that the number of effectively independent draws from the posterior distribution from of all MCMC runs was adequate. Model selection was performed using marginal log-likelihoods, which represent the probability of the data given a specific model integrated over all possible parameter values[37]. Bayes factors were computed and used to estimate and compare the probabilities of the unequal and equal models given the data for each developmental character. Stepping-stone sampling (50 MCMC runs in RevBayes[37]) was used to approximate the marginal log-likelihoods[62]. The unequal model was found to be the model that best fit the data for all developmental characters (Extended Data Table 2). Nonetheless, the equal model also gives posterior probabilities similar to those of the unequal model (Extended Data Table 2), indicating that the reconstruction obtained for each dataset is robust to the evolutionary model assumed. Finally, we assessed the sensitivity of these posterior probabilities to the branch lengths obtained from the literature by repeating all of the above analyses, but setting all branch lengths equal to 1. The posterior probabilities obtained with all branch lengths equal to 1 were similar to those obtained from the literature (Supplementary Table 2), indicating that the reconstruction obtained for each dataset is robust to the branch lengths used.

Microscopy

We used a Zeiss Discovery V12 stereomicroscope and Zeiss Axiovision software to image embryos and ovaries. For high-resolution imaging, we used Leica SP8 confocal microscope. ImageJ2 was used for analysis of images[63].

Statistics and reproducibility

For a given gene, in situ hybridization and immunohistochemistry, the sample size for *C. floridanus* consisted of at least 30 embryos or ovarioles of similar stages; for other species that produce far fewer embryos, the sample size consisted of at least 5 embryos of similar stages. One hundred per cent of the embryos sampled showed the same expression patterns. In situ hybridization and immunohistochemistry experiments for *C. floridanus* were repeated at least eight times independently. For other species, these experiments were repeated at least four times. For RNAi experiments, phenotypes were considered reproducible if at least three independent replicates gave the same results. For qPCR, statistical analysis was performed using Graphpad Prism v7. or Microsoft Excel. Relative quantification values for *YFP* RNAi,s *abdA* RNAi and *Ubx* RNAi were calculated by the same method to ensure consistency between plotted results on the graph and for analysis of variance (ANOVA). Two-way ANOVA with replication was performed, in which *Ubx* RNAi was compared with *YFP* RNAi and *abdA* RNAi with *YFP* RNAi, treating RNAi as fixed and nine target genes as random effects. Each of the tissues (zone 1, zone2 and zone 3 + zone 4) was analysed by ANOVA as a separate experiment. The qPCR experiments were performed blind at the Genomic Platform facility at the Institute for Research in Immunology and Cancer. Fisher's exact test was performed to determine: (i) whether there is a significant difference in phenotype frequency (wild-type-like versus mild or severe) between control embryos collected from rifampicin-treated colonies versus tranplanted embryos collected from rifampicin-treated colonies. Analyses were considered statistically significant at $\alpha < 0.05$. For blinding and reproducibility, two different researchers independently performed the following steps without communicating each step: sample collection from colonies, randomization of embryos between treatments, treatment of samples, replicate maintenance, and data acquisition and analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All relevant data are included in the Article, Extended Data and Supplementary Information. Raw sequence data that support the findings of this study have been deposited in GenBank with accession code MH801205, and in NCBI Sequence Read Archive with the accession code PRJNA625680. All raw image data that support the findings of this study are available in FigShare with the following identifiers: reference number 78072 (https://figshare.com/projects/The_origin_and_elaboration_of_a_major_evolutionary_transition_in_ants/78072); Fig. 1, https://doi.org/10.6084/m9.figshare.12133308; Fig. 2, https://doi.org/10.6084/m9.figshare.12133311; Fig. 3, https://doi.org/10.6084/m9.figshare.12133314; Fig. 4, https://doi.org/10.6084/m9.figshare.12133326; Extended Data Fig. 1, https://doi.org/10.6084/m9.figshare.12133296; Extended Data Fig. 2, https://doi.org/10.6084/m9.figshare.12133287; Extended Data Fig. 3, https://doi.org/10.6084/m9.figshare.12133110; Extended Data Fig. 4, https://doi.org/10.6084/m9.figshare.12133278; Extended Data Fig. 5, https://doi.org/10.6084/m9.figshare.12130902; Extended Data Fig. 6, https://doi.org/10.6084/m9.figshare.12131022; Extended Data Fig. 7, https://doi.org/10.6084/m9.figshare.12132993; Extended Data Fig. 8, https://doi.org/10.6084/m9.figshare.12131430. Source data are provided with this paper.

39. Bhatkar, A. & Whitcomb, W. Artificial diet for rearing various species of ants. *Fla. Entomol.* **53**, 229–232 (1970).
40. Khila, A. & Abouheif, E. In situ hybridization on ant ovaries and embryos. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5250 (2009).
41. Rafiqi, A. M., Lemke, S. & Schmidt-Ott, U. *Megaselia abdita*: fixing and devitellinizing embryos. *Cold Spring Harb. Protoc.* **2011**, pdb.prot5602 (2011).
42. Rothwell, W. F. & Sullivan, W. in *Drosophila* Protocols (eds Sullivan, W. et al.) 141–157 (Cold Spring Harbor Laboratory Press, 2000).
43. Bownes, M. A photographic study of development in the living embryo of *Drosophila melanogaster. J. Embryol. Exp. Morpol.* **33**, 789–801 (1975).
44. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
45. Bonasio, R. et al. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**, 1068–1071 (2010).
46. Kosman, D. et al. Multiplex detection of RNA expression in *Drosophila* embryos. *Science* **305**, 846 (2004).
47. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 1989).
48. Rafiqi, A. M., Lemke, S. & Schmidt-Ott, U. *Megaselia abdita*: preparing embryos for injection. *Cold Spring Harb. Protoc.* **2011**, pdb.prot5601 (2011).
49. Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl Acad. Sci. USA* **88**, 7276–7280 (1991).
50. Xie, F., Xiao, P., Chen, D., Xu, L. & Zhang, B. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol. Biol.* (2012).
51. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.0031 (2002).
52. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 33 (2006).
53. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper—Excel-based tool using pair-wise correlations. *Biotechnol. Lett.* **26**, 509–515 (2004).
54. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250 (2004).
55. Borowiec, M. L. et al. Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. *Mol. Phylogenet. Evol.* **134**, 111–121 (2019).
56. Blaimer, B. B. et al. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evol. Biol.* **15**, 271 (2015).
57. Mezger, D. & Moreau, C. S. Out of South-East Asia: phylogeny and biogeography of the spiny ant genus *Polyrhachis* Smith (Hymenoptera: Formicidae). *Syst. Entomol.* **41**, 369–378 (2016).
58. Lilienstern, M. Beiträge zur Bakteriensymbiose der Ameisen. *Zeitschrift für Morphologie und Ökologie der Tiere* **26**, 110–134 (1932).
59. Jungen, H. Endosymbionten bei Ameisen. *Insectes Soc.* **15**, 227–232 (1968).
60. Pagel, M., Meade, A. & Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53**, 673–684 (2004).
61. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

62. Fan, Y., Wu, R., Chen, M.-H., Kuo, L. & Lewis, P. O. Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.* **28**, 523–532 (2011).
63. Rueden, C. T. et al. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017).

Acknowledgements We thank L. Davis, R. Johnson, A. Suarez, J. Gibson, J. Rand, A. Wild and E. LeBrun for help with collecting ants; A. Vasquez-Correa, S. Joly, P. Ward and T. Oakley for help with ancestral-state reconstruction; M. Zayd and T. Chen for help with qPCR analysis; C. Metzl for translations; P. Lasko, S. F. Gilbert, J. Liebig, D. W. Wheeler, R. Rajakumar, C. Extavour, Y. Idaghdour, D. Schoen, A. Khila and members of the laboratory of E.A. for discussions or comments on the manuscript; and McGill University's Integrated Quantitative Biology Initiative (IQBI) and Advanced BioImaging Facility (ABIF) for imaging support. This work was supported by a doctoral fellowship from FQRNT (Quebec) to A.R., a Bezmialem Vakif University Grant (Turkey) to A.M.R., and an NSERC Discovery Grant and Steacie Fellowship (Canada), John Simon Guggenheim Fellowship (USA) and KLI Fellowship (Austria) to E.A.

Author contributions E.A., A.M.R. and A.R. conceived the project, designed experiments and collected ants. A.R. and A.M.R. performed all experiments. E.A. performed phylogenetic analyses. E.A., A.M.R. and A.R. wrote the manuscript.

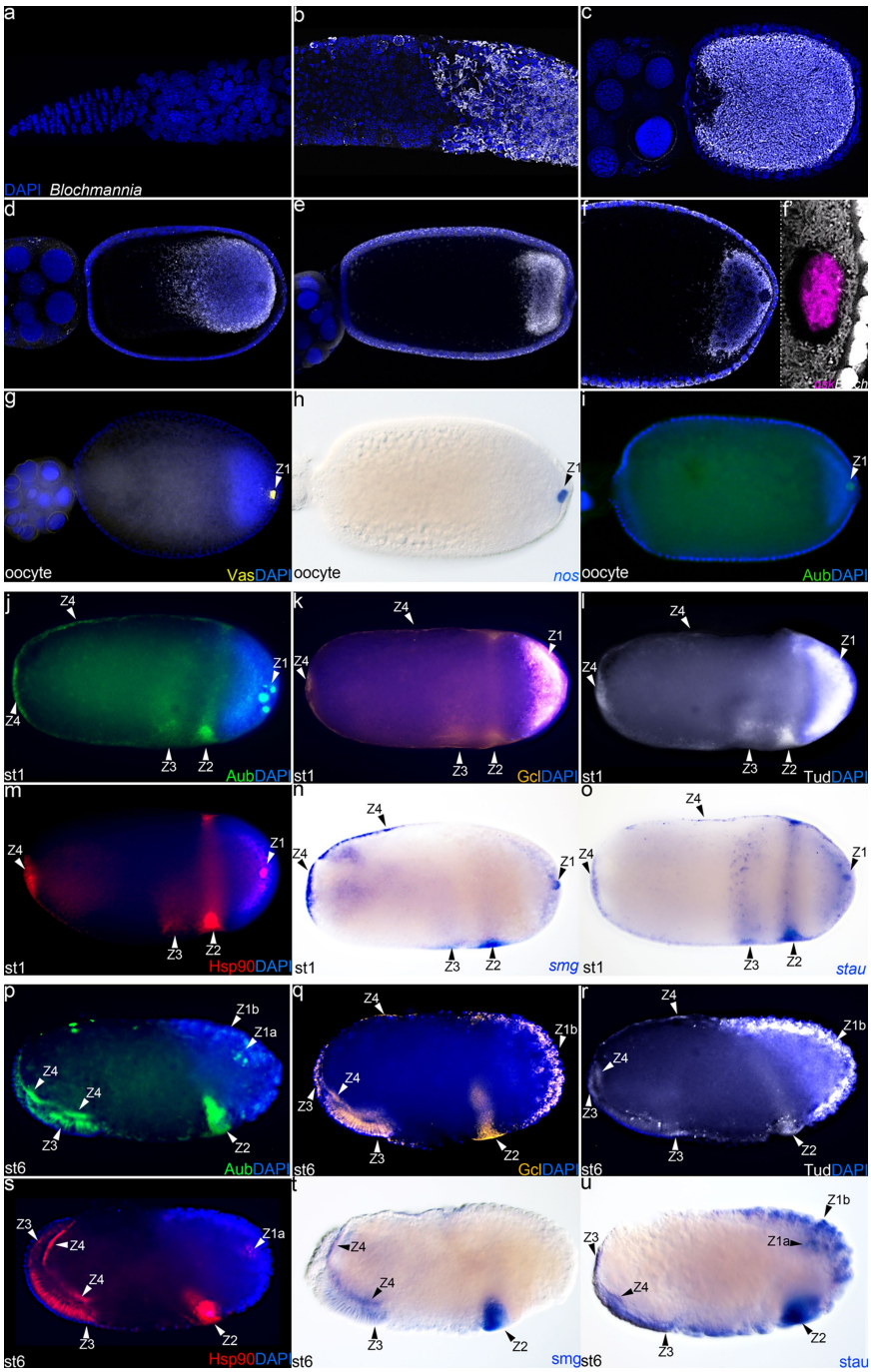
Competing interests The authors declare no competing interests.

Additional information

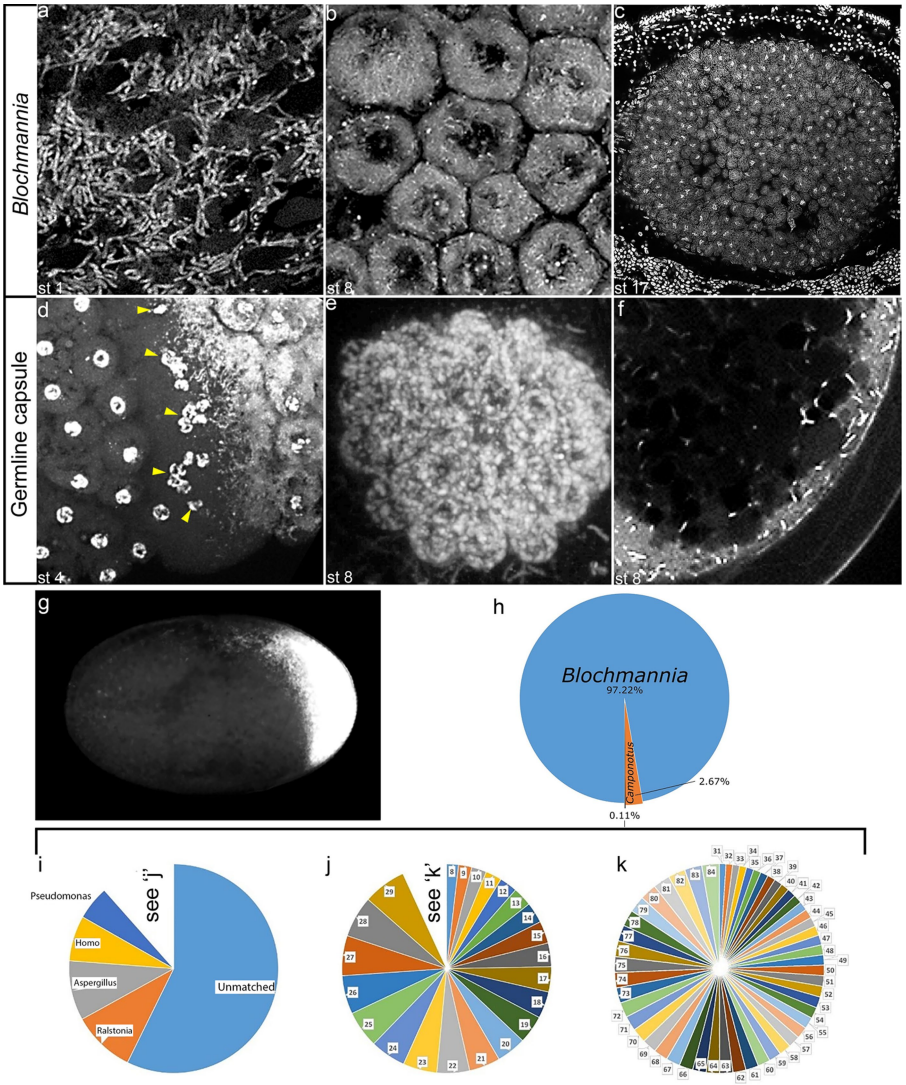
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2653-6.

Correspondence and requests for materials should be addressed to E.A.

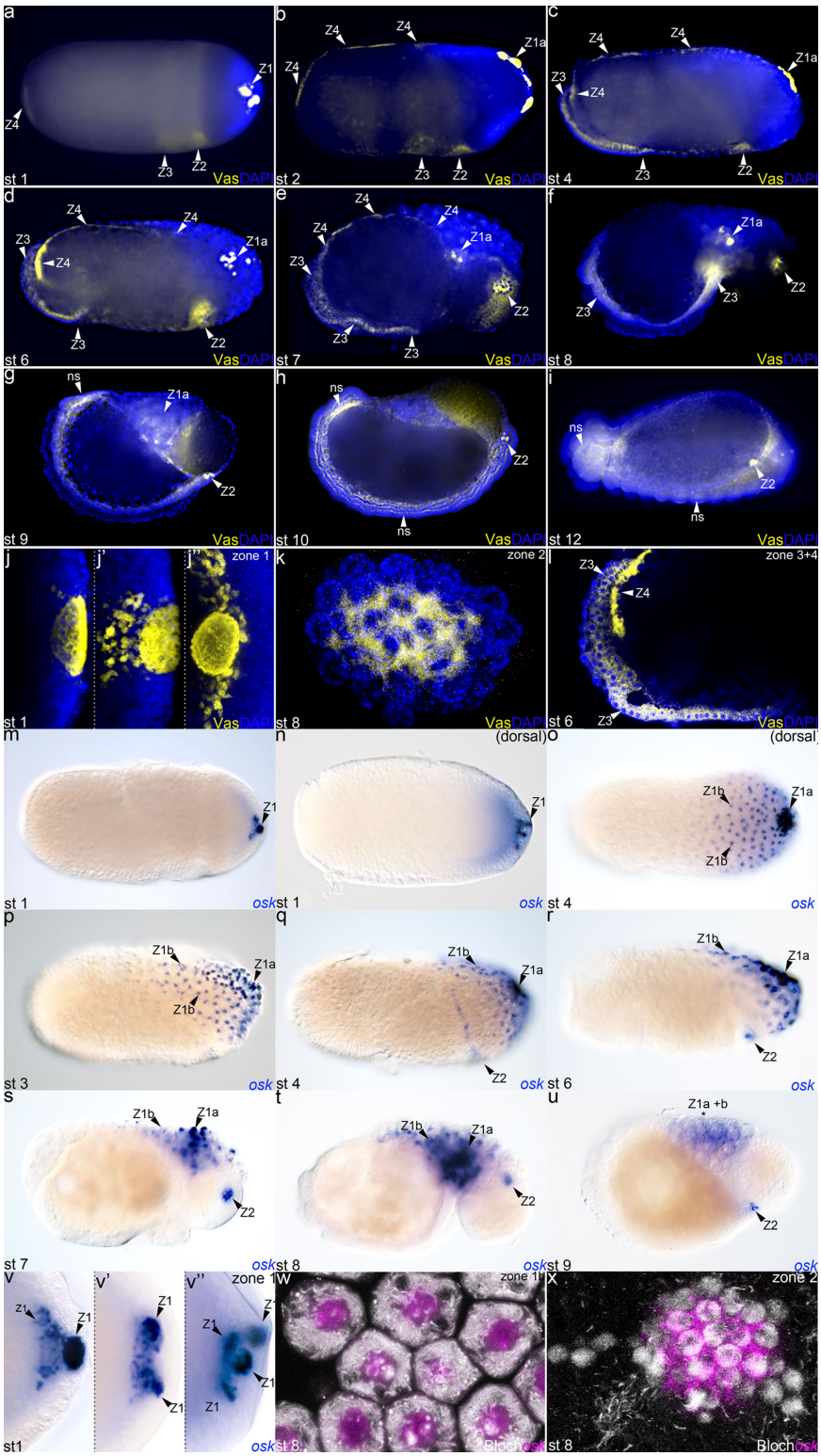
Peer review information *Nature* thanks Cameron R. Currie, Yannick Wurm and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Distribution of *Blochmannia* during oogenesis and the subcellular localization and expression of germline genes in *C. floridanus* oocytes and embryos. a–**f**, **f'**, Ovaries showing nuclear-stain DAPI in blue and *Blochmannia* in white: germline stem-cell niche without *Blochmannia* (**a**), germarium in which *Blochmannia* colonization occurs (**b**), *Blochmannia* initially fill the entirety of the cytoplasm of young oocytes (**c**) and progressively localize to the posterior pole of older oocytes (**d**–**f**), where *Blochmannia* surrounds the germplasm (**f'**). **f'**, **g**–**i**, Mature oocytes showing maternal expression of germline genes in oocytes, showing *osk* mRNA in magenta (**f'**), Vas protein in yellow (**g**), *nos mRNA* in blue (**h**), Aub protein in green (**i**) and nuclear-stain DAPI in blue. **j**–**o**, Subcellular localization zones in stage (st)-1 freshly laid eggs showing Aub protein in green (**j**), Gcl protein in orange (**k**), Tud protein in white (**l**), Hsp90 protein in red (**m**), *smg* mRNA in blue (**n**) and *stau* mRNA in blue (**o**). **p**–**u**, Expression in stage-6 cellular blastoderm embryos showing Aub protein in green (**p**), Gcl protein in orange (**q**), Tud protein in white (**r**), Hsp90 protein in red (**s**), *smg* mRNA in blue (**t**) and *stau* mRNA in blue (**u**). Arrowheads indicate subcellular localization or expression zones of germline genes: zone 1, zone 1a, zone 1b, zone 2, zone 3 and zone 4. Anterior is to the left, dorsal is to the top. In situ hybridization and immunohistochemistry experiments were repeated at least 8 times independently on $n \geq 30$ oocytes or embryos per developmental stage.

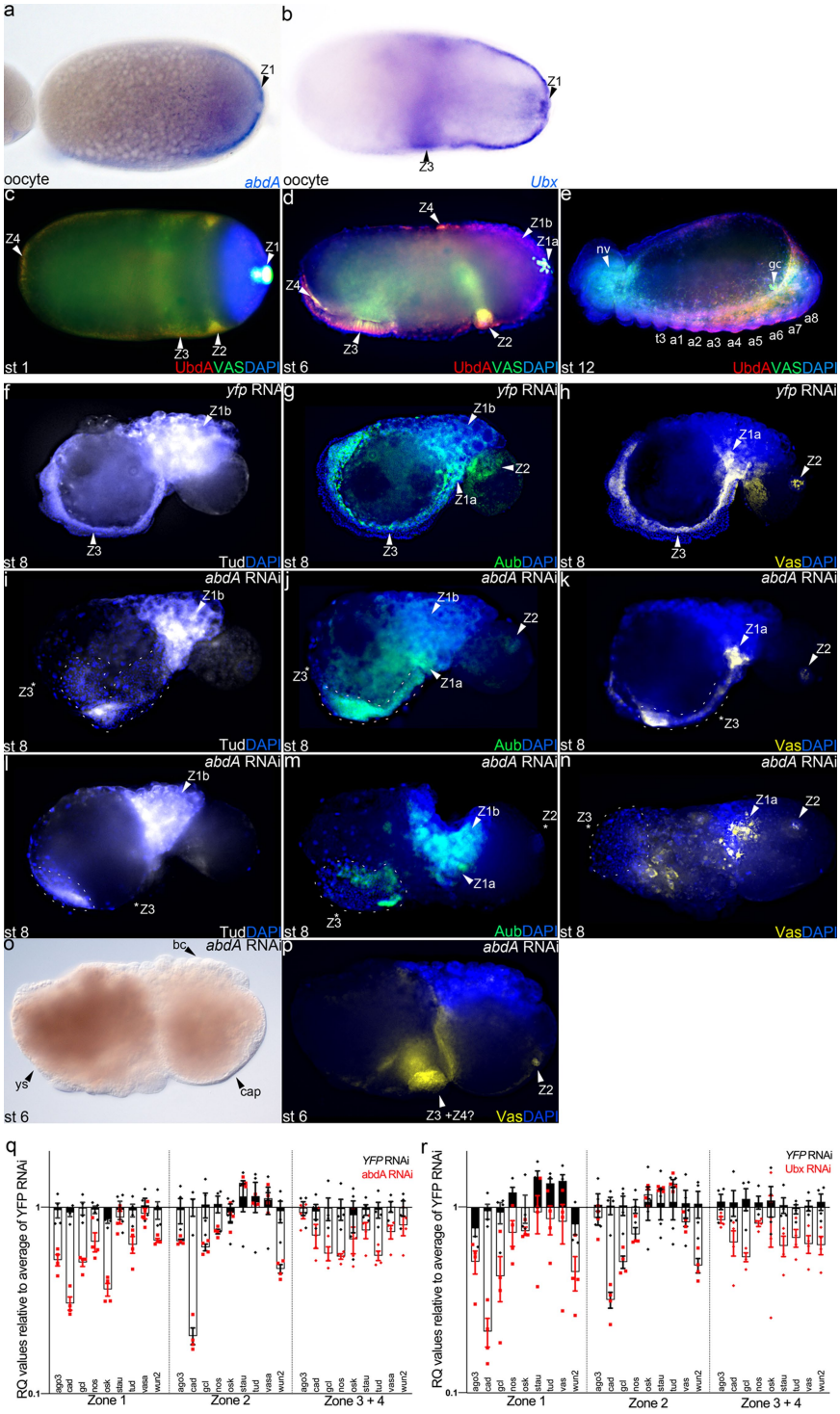


Extended Data Fig. 2 | *Blochmannia* segregates between bacteriocytes and germline capsule, and makes up 97.2% of DNA content in freshly laid eggs of *C. floridanus*. **a**–**f**, *Blochmannia*: at the posterior pole in freshly laid stage-1 eggs (**a**); inside bacteriocytes in stage-8 embryos (**b**); in bacteriocytes that line the midgut of stage-17 embryos (**c**); together with germline-precursor nuclei (yellow arrowheads) along the crest of the future germline capsule (**d**); surrounding the novel germline within the germline capsule (**e**); and as a small seed population for vertical transmission in the germline capsule (**f**). **g**, Freshly laid egg with DAPI in white, showing few zygotic nuclei in the anterior and *Blochmannia* at the posterior pole. **h**–**k**, Pie charts representing the number of Illumina Hi-Seq reads that match each of the indicated genera from DNA of freshly laid eggs. **h**, High abundance of *Blochmannia* DNA (blue) compared to that of host DNA (orange) and of other associated microorganisms (slim black slice) (shown in more detail in **i**–**k**) of decreasing abundance. We used a sequence similarity (*e*-value) of e^{-3} as a cut-off value for including any genus in our analysis. Numbers in **j** and **k** represent the following species: 8, *Serratia*; 9, *Leuconostoc*; 10, *Cupriavidus*; 11, *Cutibacterium*; 12, *Corynebacterium*; 13, *Mycobacterium*;14, *Candida*; 15, *Cyberlindnera*; 16, *Lactobacillus*; 17, *Brevibacterium*; 18, *Methylobacterium*; 19, *Pan*; 20, *Staphylococcus*; 21, *Sphingomonas*; 22, *Bradyrhizobium*; 23, *Plasmopara*; 24, *Bacillus*; 25, *Streptococcus*; 26, *Sphingopyxis*; 27, *Hyphomicrobium*, 28, *Acinetobacter*; 29, uncultured; 30, see **k**; 31, *Burkholderia*; 32, *Achromobacter*; 33, *Pichia*; 34, *Hyphopichia*; 35, *Penicillium*; 36, *Cyprinus*; 37, *Paenibacillus*; 38, *Brachybacterium*; 39, *Stenotrophomona*; 40, *Variovorax*; 41, *Streptomyces*; 42, *Sphingobium*; 43, *Nocardiopsis*; 44, *Dermabacter*; 45, *Sphingobacteriu*; 46, *Klebsiella*; 47, *Morganella*; 48, *Acidovorax*; 49, *Malassezia*; 50, *Lysobacter*; 51, *Rothia*; 52, *Pongo*; 53, *Rhodoplanes*; 54, *Microbacterium*; 55, *Rhodopseudomona*; 56, *Acheta*; 57, *Exiguobacterium*; 58, *Paraburkholderi*; 59, *Enterococcus*; 60, *Ramlibacter*; 61, *Actinomyces*; 62, *Bordetella*; 63, *Xanthomonas*; 64, *Brevundimonas*; 65, *Citrobacter*; 66, *Drosophila*; 67, *Lactococcus*; 68, *Mesorhizobium*; 69, *Candidatus*; 70, *Gluconobacter*; 71, *Rhodococcus*; 72, *Rubrivivax*; 73, *Saccharomyces*; 74, *Chelatococcus*; 75, *Hydrogenophaga*; 76, *Micrococcus*; 77, *Rhizobium*; 78, *Thauera*; 79, *Azospirillum*; 80, *Bosea*; 81, *Micromonospora*; 82, *Caulobacter*; 83, *Triticum*; 84, *Tsukamurella*. DAPI staining was repeated at least 4 times on $n \geq 30$ embryos per developmental stage.



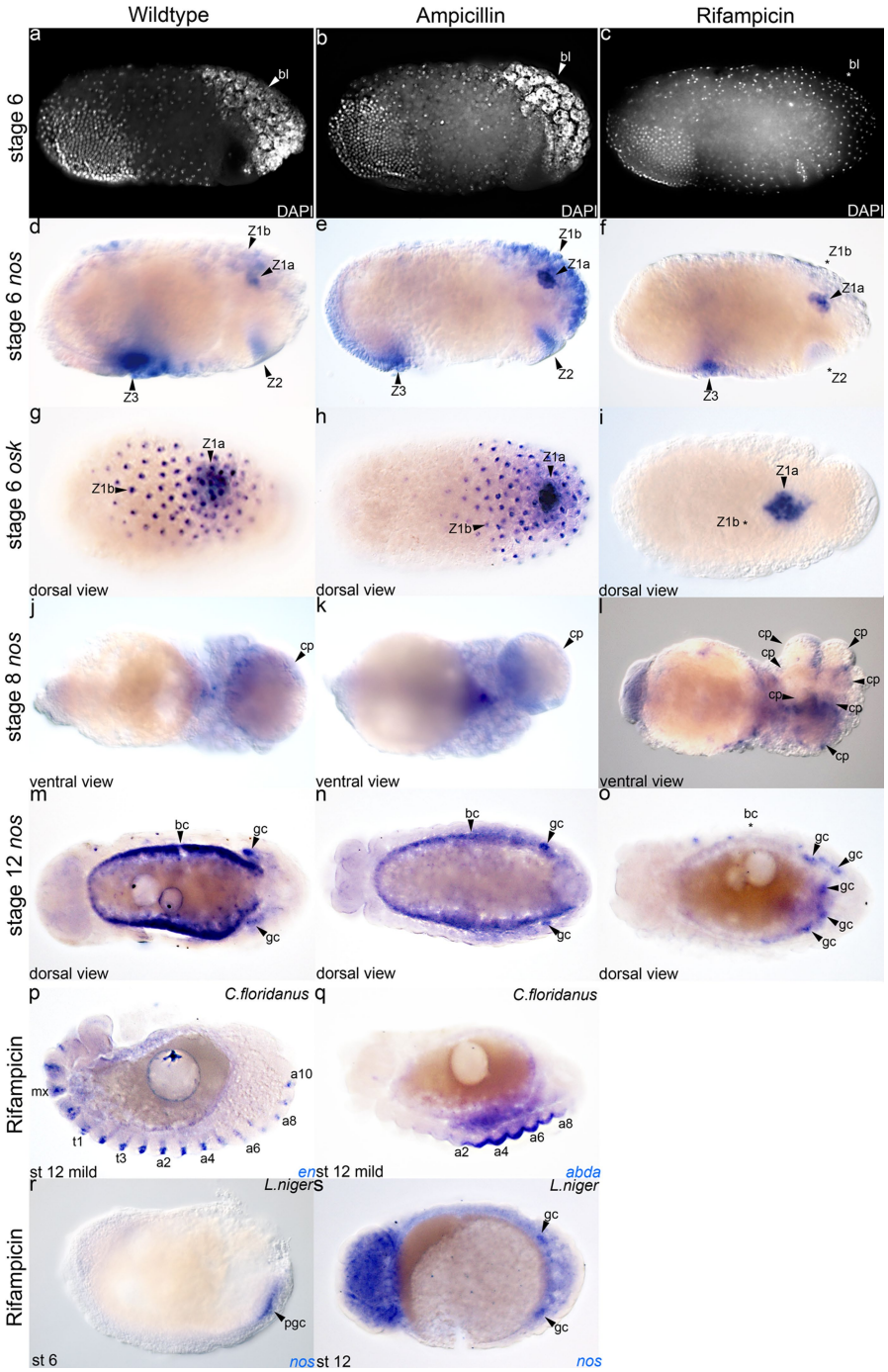
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Tracking the four functionally distinct zones through *C. floridanus* embryogenesis. a–**i**, Embryos showing Vas protein staining in yellow and DAPI in blue: freshly laid stage-1 and -2 eggs (**a**, **b**), cellular blastoderm stage-4 and -6 (**c**, **d**), gastrulation stage-7 (**e**), germband extension stage-8 to -10 (**f**–**h**) and segmentation stage-12 (**i**) embryos. **j**, **j′**, **j″**, **k**, **l**, Embryos showing higher-magnification confocal images of zone 1–4: freshly laid stage-1 egg, showing small germplasm foci budding off of the ancestral germplasm (**j**, **j′**, **j″**), stage-8 embryo showing novel germline (**k**), stage-6 embryo showing germband (zone 3) and yolk sac (zone 4) expression (**l**). ns, onset of Vas expression throughout the nervous system, brain and central nervous system in embryos from stage 9 onwards. **m**–**u**, *osk* mRNA in blue in stage-1 freshly laid egg (**m**, **n**), cellular blastoderm stage-3 and -4 embryo (**o**–**r**), gastrulation stage-7 embryo (**s**), and germband extension stage-8 and -9 embryo (**t**, **u**). **n**, **o**, Dorsal view, showing localization of small germplasm foci within the centre of bacteriocytes (zone 1b). **q**–**u**, Formation of the novel germline (zone 2). **u**, Embryo, showing loss of zone 1a and zone 1b. **v**, **v′**, **v″**, Small foci budding off the ancestral germplasm (zone 1). **w**, **x**, Higher-magnification confocal images of embryos, showing *osk* mRNA in magenta and DAPI in white. **w**, Stage-8 embryo, showing *osk* mRNA in magenta in the centre of bacteriocytes (zone 1b) surrounded by bacteria. **x**, Stage-8 embryo, showing expression of *osk* mRNA in the novel germline (zone 2). Zones are indicated with arrowheads. Anterior is to the left, dorsal is to the top. In situ hybridization and immunohistochemistry experiments were repeated at least 8 times independently on $n \geq 30$ embryos per developmental stage.

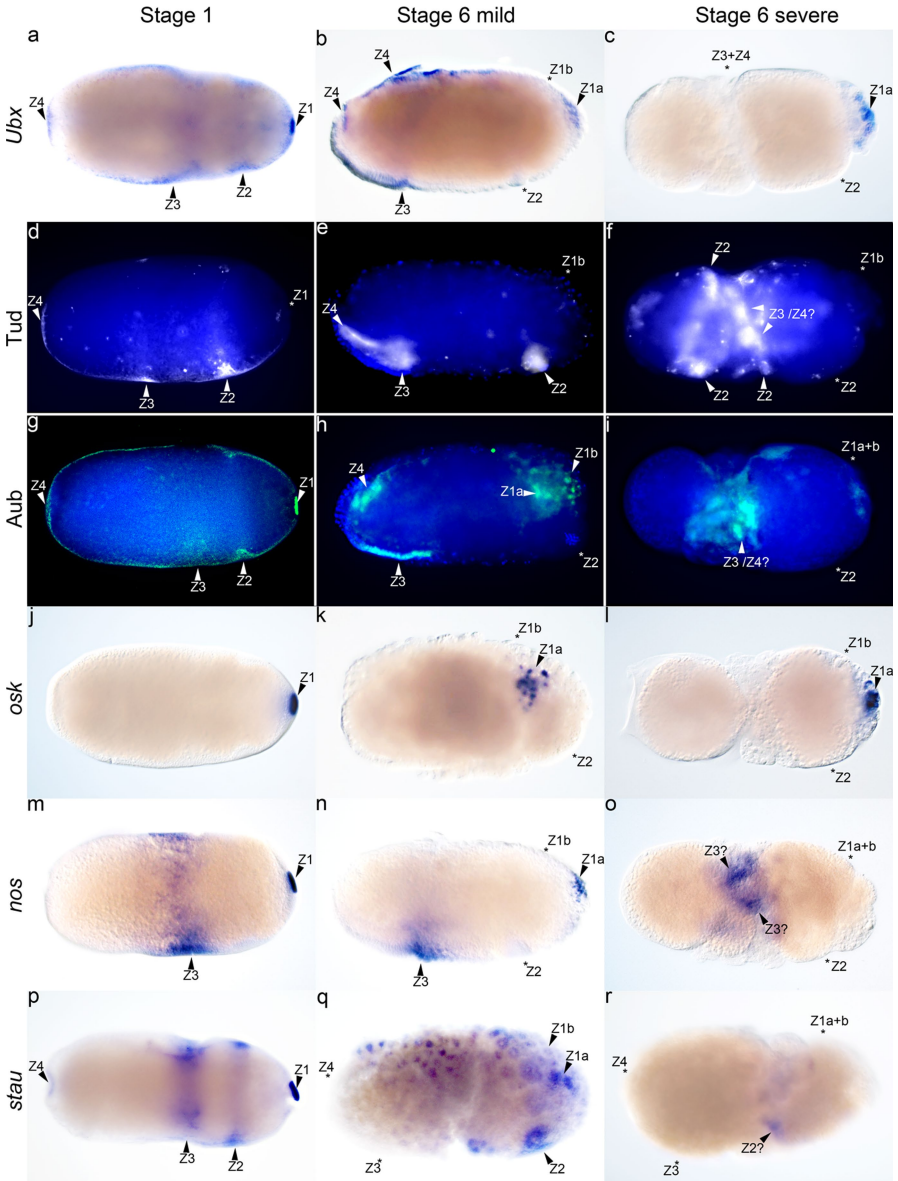


Extended Data Fig. 4 | See next page for caption.

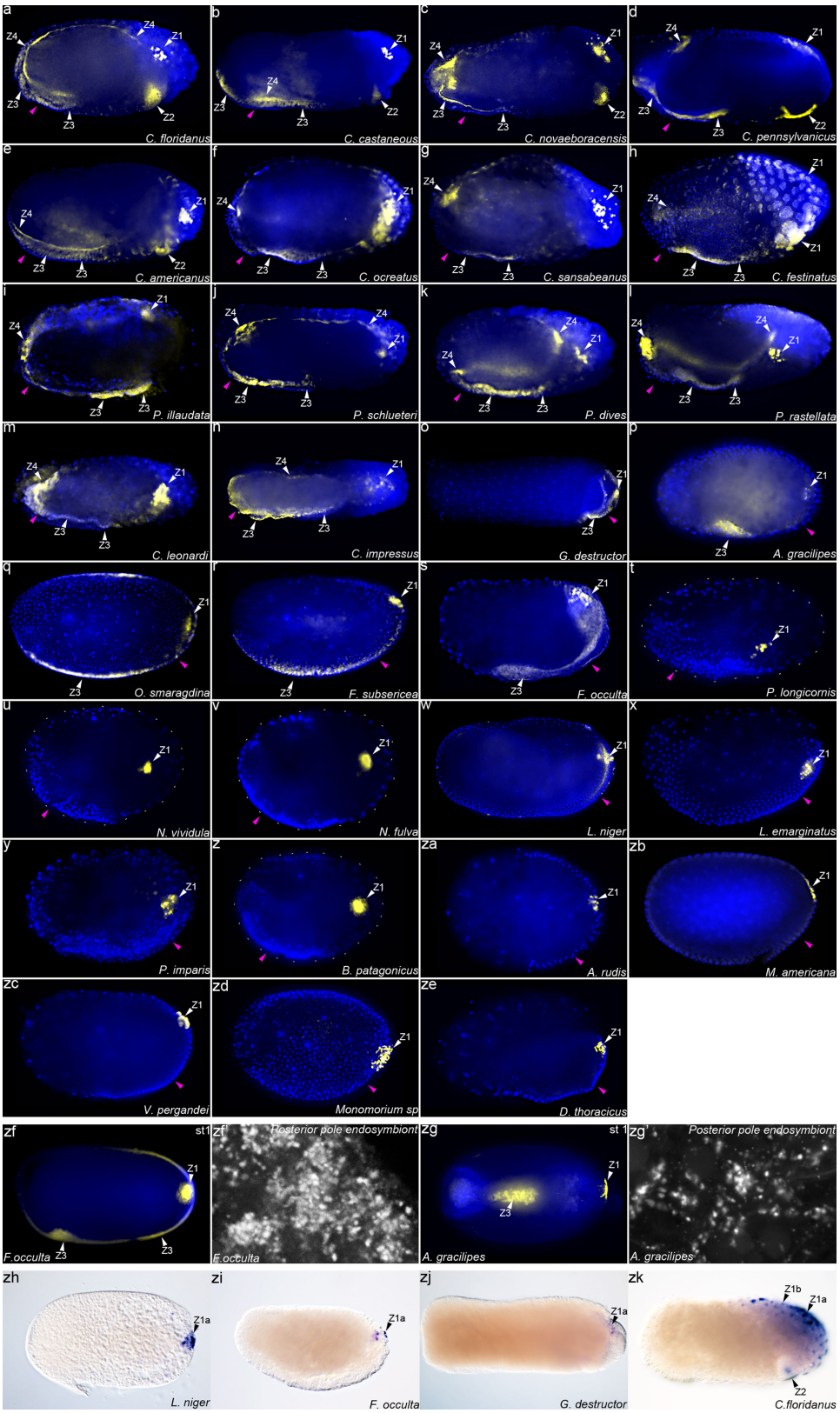
Extended Data Fig. 4 | *abdA* and *Ubx* are upstream of the germline genes. **a**, **b**, Mature oocytes stained for *abdA* mRNA (**a**) or *Ubx* mRNA (**b**) in blue. **c**–**e**, Colocalization (yellow and orange) of Ubx and AbdA (UbdA) protein in red, Vas protein in green and DAPI in blue in freshly laid stage-1 eggs (**c**), and stage-6 (**d**) and stage-12 (**e**) wild-type embryos. **f**–**p**, Expression of the germline genes in *YFP* RNAi ($n = 81$) (**f**–**h**) and high-concentration *abdA* RNAi embryos ($n = 61$ out 69) with DAPI in blue (**j**–**p**), stained for Tud in white (**f**, **i**, **l**), Aub in green (**g**, **j**, **m**) or Vas in yellow (**h**, **k**, **n**, **p**), and DIC of stage-6 embryo with severe phenotype (**o**). **i**–**k**, *abdA* RNAi embryos that are split along the midline ($n = 21$ out 61). **l**–**p**, Severe *abdA* RNAi phenotypes with an undifferentiated stub ($n = 34$ out of 61) (**l**–**n**) or in which the embryo is not detectable (**o**, **p**) ($n = 6$ out 61). Dotted outlines show changes in germband morphology and zone-3 expression after *abdA* RNAi. Zones are indicated with arrowheads. Asterisks indicate loss of germline gene expression within a specific zone. bc, bacteriocytes; cap, giant capsule; ys, yolk sac. Anterior is to the left, dorsal is to the top. **q**, **r**, Tissue-specific qPCR of nine germline genes (*x* axis; *ago3*, *cad*, *gcl*, *nos*, *osk*, *stau*, *tud*, *vasa* and *wun2*) from zone 1 (bacteriocytes), zone 2 (germline capsules), and zone 3 + zone 4 (embryonic germband + yolk sac) following *YFP* RNAi, low-concentration *abdA* RNAi and *Ubx* RNAi. Open bars represent mean relative quantification (RQ) values (*x* axis) and error bars represent standard error of the mean of: *abdA* RNAi (**q**) or *Ubx* RNAi (**r**). Black bars represent mean relative quantification values (*y* axis) and error bars represent standard error of the mean of *YFP* RNAi controls. Each individual data point (red squares) represents relative quantification value of a technical replicate from *abdA* or *Ubx* RNAi treatment relative to the average of all replicates of *YFP* RNAi control treatments (black diamonds) in that tissue. Two-tailed two-way ANOVA with replication for *abdA* RNAi versus YFP RNAi in zone 1 ($F = 129.311$, degrees of freedom (d.f.) = 1, $n = 54$, $P = 5.95504 \times 10^{-16}$ for *abdA* RNAi); zone 2 ($F = 20.733$, d.f. = 1, $n = 54$, $P = 3.04542 \times 10^{-5}$ for *abdA* RNAi); zone 3 + zone 4 ($F = 38.932$, d.f. = 1, $n = 54$, $P = 7.02605 \times 10^{-8}$ for *abdA* RNAi). Two-tailed two-way ANOVA with replication for *Ubx* RNAi versus YFP RNAi in zone 1 ($F = 66.278$, d.f. = 1, $n = 54$, $P = 5.84252 \times 10^{-11}$ for *Ubx* RNAi); zone 2 ($F = 12.628$, d.f. = 1, $n = 54$, $P = 0.000798519$ for *Ubx* RNAi); zone 3 + zone 4 ($F = 40.841$, d.f. = 1, $n = 54$, $P = 4.00577 \times 10^{-8}$ for *Ubx* RNAi). Raw data are in Source Data. In situ hybridization and immunohistochemistry experiments (**a**–**e**) were repeated at least 8 times independently on $n \geq 30$ embryos per developmental stage.



Extended Data Fig. 5 | Antibiotic treatment does not show unspecific effects. a–**c**, Stage-6 cellular blastoderm *C. floridanus* embryos with DAPI in white from wild-type colonies ($n \geq 30$ embryos) (**a**), colonies treated with ampicillin ($n \geq 15$ embryos) (**b**) or colonies treated with rifampicin ($n \geq 15$ embryos) (**c**). **d**–**o**, *C. floridanus* embryos stained for *nos* mRNA (**d**–**f**, **j**–**o**) and *osk* mRNA (**g**–**i**) in blue collected from wild-type colonies ($n \geq 30$ embryos each) (**d**, **g**, **j**, **m**), colonies treated with ampicillin ($n \geq 15$ embryos each) (**e**, **h**, **k**, **n**) or colonies treated with rifampicin ($n \geq 15$ embryos each) (**f**, **i**, **l**, **o**). **p, q**, Stage-12 mild-phenotype embryos collected from rifampicin-treated *C.-floridanus* colonies, showing expression of the segment polarity gene *en* in blue ($n \geq 15$ embryos) (**p**) or *abdA* mRNA in blue ($n \geq 15$ embryos) (**q**). **r**, **s**, *Lasius niger* embryos collected from rifampicin-treated colonies showing *nos* mRNA in blue in stage-6 embryos with normal primordial germ cells (pgc) ($n \geq 5$ embryos) (**r**) and stage-12 embryos with normal germ cells (gc) ($n \geq 5$ embryos) (**s**). Segments marked are as following: maxillary (mx), thoracic segments 1–3 (t1–t3) and abdominal segments 1–10 (a1–a10). White arrowheads indicate presence of *Blochmannia* (bl). White and black asterisks in embryos from rifampicin-treated colonies indicate loss of *Blochmannia* or loss of germline gene expression. **d**–**i**, Black arrowheads indicate zones. **j**–**l**, Black arrowheads indicate germline capsule(s) (cp). **m**–**o**, Black arrows indicate normal bacteriocyte (bc) and gonads (gc), development. Anterior is to the left, dorsal is to the top (**a**–**f**, **p**-**r**); dorsal is towards the reader in **g**–**i**, **m**–**o**, **s**; and ventral is towards the reader in **j**–**l**. In situ hybridization experiments were repeated at least 8 times (*C. floridanus*) or 4 times (*L. niger*) independently.

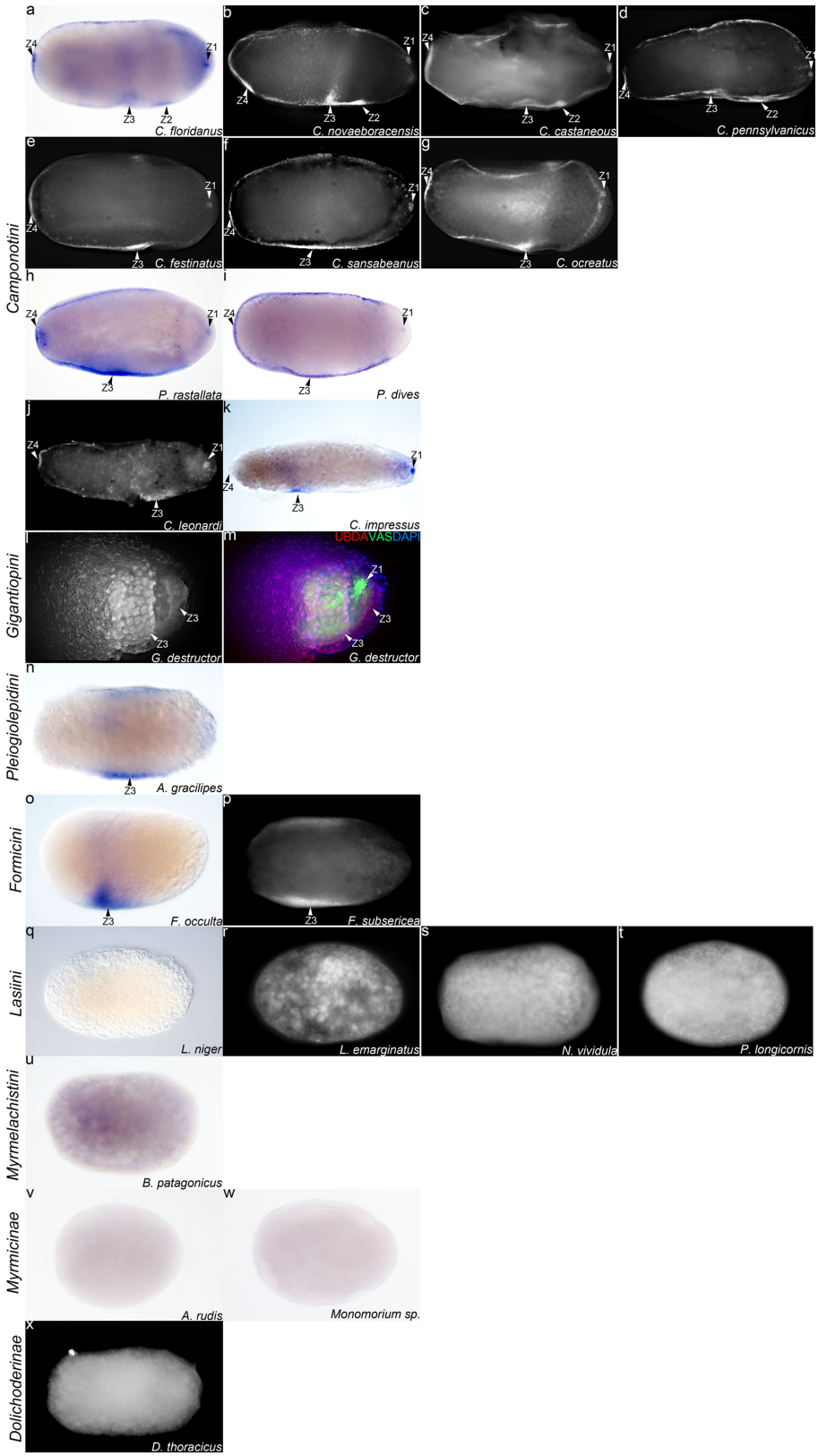


Extended Data Fig. 6 | *Blochmannia* maintains and selectively regulates mRNAs and proteins of maternal Hox and germline genes. a–**r**, Embryos from rifampicin-treated colonies stained for *Ubx* mRNA in blue (**a**–**c**), Tud protein in white (**d**–**f**), Aub protein in green (**g**–**i**), *osk* mRNA in blue (**j**–**l**), *nos* mRNA in blue (**m**–**o**) or *stau* mRNA in blue (**p**–**r**). **a**, **d**, **g**, **j**, **m**, **p**, Freshly laid stage-1 eggs showing no effect on the number of zones relative to wild type, except for in **d** Tud in white showing loss of zone 1 relative to wild type. **b**, **e**, **h**, **k**, **n**, **q**, Stage-6 mild-phenotype embryos with no observable morphological defects: asterisks indicate loss of specific mRNAs and proteins of Ubx and germline gene expression. **c**, **f**, **i**, **l**, **o**, **r**, Stage-6 severe-phenotype embryos showing morphological defects and loss or misexpression of germline and Hox genes. **d**–**i**, Fluorescent images with DAPI in blue. Zones of germline and Hox gene expression are indicated with arrowheads. Question marks indicate presumptive zones. Anterior is to the left, dorsal is to the top. In situ hybridization and immunohistochemistry experiments were repeated at least 4 times independently on $n \geq 15$ embryos per developmental stage.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Character states of germline localization zones, location of embryo, obligate endosymbiont and germline capsule. **a**–**z**, **za**, **zb**, **zc**, **zd**, **ze**, Cellular blastoderm stage embryos from Formicinae (**a**–**z**) and two sister subfamilies (**za**, **zb**, **zc**, **zd**) Myrmicinae and Dolichoderinae (**ze**), stained for Vas protein in yellow with DAPI in blue. **a**–**n**, Camponotini tribe. **a**, *Camponotus floridanus*. **b**, *Camponotus castaneous*. **c**, *Camponotus novaeboracensis*. **d**, *Camponotus pennsylvanicus*. **e**, *Camponotus americanus*. **f**, *Camponotus ocreatus*. **g**, *Camponotus sansabeanus*. **h**, *Camponotus festinatus*. **i**, *Polyrhachis illaudata*. **j**, *Polyrhachis schlueteri*. **k**, *Polyrhachis dives*. **l**, *Polyrhachis rastallata*. **m**, *Colobopsis leonardi*. **n**, *Colobopsis impressus*. **o**, Gigantiopini tribe: *Gigantiops destructor*. **p**, Pleigiolepidini tribe: *Anoplolepis gracilipes*. **q**, Oecophyllini tribe: *Oecophylla smaragdina*. **r**, **s**, Formicini tribe. **r**, *Formica subsericea*. **s**, *Formica occulta*. **t**–**y**, Lasiini tribe. **t**, *Paratrechina longicornis*. **u**, *Nylanderia vividula*. **v**, *Nylanderia fulva*. **w**, *Lasius niger*. **x**, *Lasius emarginatus*. **y**, *Prenolepis imparis*. **z**, Myrmelachistini tribe: *Brachymyrmex patagonicus*. **za**, **zb**, **zc**, **zd**, Myrmicinae. **za**, *Aphaenogaster rudis*. **zb**, *Myrmica americana*. **zc**, *Veromessor pergandei*. **zd**, *Monomorium* sp. **ze**, Dolichoderinae: *Dolichoderus thoracicus*. **zf**, **zg**, Freshly laid stage-1 eggs stained for Vas protein in yellow with DAPI in blue of *F. occulta* (**zf**) and *A. gracilipes* (**zg**). **zf′**, **zg′**, Endosymbiont at the posterior pole of *F. occulta* (**zf′**) and *A. gracilipes* (**zg′**). **zi**, **zj**, **zk**, **zl**, Cellular blastoderm stage embryos showing *osk* mRNA in blue, for *L. niger* (**zi**), *F. occulta* (**zj**), *G. destructor* (**zk**) and *C. floridanus* (**zl**). Zones of germline gene expression are indicated with white or black arrowheads. Magenta arrowheads indicate the location of the embryo within the egg. Experiments on all species were repeated 4 times independently on $n \geq 5$ embryos, except for *C. floridanus*, which was repeated 8 times independently with $n = 30$. Anterior is to the left, dorsal is to the top.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Character states of maternal Hox localization zones. **a**–**j**, **l**–**x**, Freshly laid stage-1 eggs from the Formicinae (**a**–**u**) and two sister subfamilies Myrmicinae (**v**, **w**) and Dolichoderinae (**x**) stained for UbdA (*Ubx* + *abdA* protein) in white or blue and (in **k**) *abdA* mRNA in blue. **a**–**k**, Camponotini tribe. **a**, *Camponotus floridanus*. **b**, *Camponotus novaeboracensis*. **c**, *Camponotus castaneous*. **d**, *Camponotus pennsylvanicus*. **e**, *Camponotus festinatus*. **f**, *Camponotus sansabeanus*. **g**, *Camponotus ocreatus*. **h**, *Polyrhachis rastallata*. **i**, *Polyrhachis dives*. **j**, *Colobopsis leonardi*. **k**, *Colobopsis impressus*. **l**, **m**, Gigantiopini tribe: *Gigantiops destructor*. In **m**, UbdA protein in red co-stained with Vas protein in green and DAPI in blue to distinguish germ cells from zone 3. **n**, *Pleigiolepidini* tribe: *Anoplolepis gracilipes*. **o**, **p**, *Formicini* tribe. **o**, *Formica occulta*. **p**, *Formica subsericea*. **q**–**t**, Lasiini tribe. **q**, *Lasius niger*. **r**, *Lasius emargiatus*. **s**, *Nylanderia vividula*. **t**, *Paratrechina longicornis*. **u**, Myrmelachistini tribe: *Brachymyrmex patagonicus*. **v**, **w**, Myrmicinae subfamily. **v**, *Aphaenogaster rudis*. **w**, *Monomorium* sp. **x**, Dolichoderinae subfamily: *Dolichoderus thoracicus*. Zones of maternal Hox localization are indicated with arrowheads: zone 1 (ancestral germline), zone 2 (novel germline), zone 3 (embryo) and zone 4 (anterior). Anterior is to the left, dorsal is to the top. Experiments on all species were repeated 4 times independently on $n \geq 5$ embryos, except for *C. floridanus*, for which experiments were repeated 8 times independently with $n = 30$ embryos.

Extended Data Table 1 | Combinatorial and dynamic localization or expression of germline and Hox genes across zones in stage-1 eggs and stage-6 embryos

		Wildtype stage 1				Wildtype stage 6				
#	Genes used	zone1	zone2	zone3	zone4	zone1a	zone1b	zone2	zone3	zone4
1	**Vasa**	+	+	+	+	+	-	+	+	+
2	***oskar***	+	-	-	-	+	+	+	-	-
3	***nanos***	+	-	+	-	+	+	+	+	-
4	**Tud**	+	+	+	+	-	+	+	+	+
5	**Aub**	+	+	+	+	+	+	+	+	+
6	***staufen***	+	+	+	+	+	+	+	+	+
7	***abdA***	+	-	+	-	+	+	+	+	+
8	***Ubx***	+	+	+	+	+	+	+	+	+
9	**Gcl**	+	+	+	+	-	+	+	+	+
10	**Hsp90**	+	+	+	+	+	-	+	+	+
11	***smg***	+	+	+	+	-	-	+	+	+

		Rifampicin stage 1				Rifampicin (mild) stage 6				
1	**Vasa**	+	+	+	+	+	-	-	+	+
2	***oskar***	+	-	-	-	+	-	-	-	-
3	***nanos***	+	-	+	-	+	-	-	+	-
4	**Tud**	-	+	+	+	-	-	+	+	+
5	**Aub**	+	+	+	+	+	+	-	+	+
6	***staufen***	+	+	+	+	+	+	+	-	-
7	***abdA***	-	-	+	-	-	-	-	+	+
8	***Ubx***	+	+	+	+	+	-	-	+	+

'+' indicates presence and '–' indicates absence. Grey shading indicates a change in localization or expression between embryos from wild-type and rifampicin-treated colonies.

Extended Data Table 2 | Posterior probabilities under the unequal- and equal-rates model for the five developmental characters

	Unequal model					Equal model				
Node #	germline localization zones	maternal Hox localization zones	obligate endo symbiont	Location of embryo	germline capsule	germline localization zones	maternal Hox localization zones	obligate endo symbiont	Location of embryo	germline capsule
0	0.95	0.91	0.95	0.96	1.00	1.00	0.96	1.00	0.92	0.98
1	0.96	0.91	0.96	0.96	1.00	1.00	0.96	1.00	0.92	0.98
2	0.97	0.92	0.97	0.97	1.00	1.00	0.95	1.00	0.95	0.98
3	0.98	0.95	0.98	0.98	1.00	1.00	0.95	1.00	0.97	0.98
4	0.99	0.95	0.99	0.99	1.00	1.00	0.93	1.00	0.99	0.98
5	0.95	0.90	0.95	0.95	1.00	1.00	0.95	1.00	0.92	0.98
6	0.89	0.85	0.94	0.95	1.00	0.95	0.89	1.00	0.92	0.98
7	0.98	0.96	0.98	0.93	1.00	1.00	0.95	1.00	0.88	0.98
8	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.99	0.98
9	1.00	0.98	1.00	0.74	1.00	1.00	0.96	1.00	0.68	0.98
10	1.00	0.98	1.00	0.61	1.00	1.00	0.96	1.00	0.62	0.98
11	1.00	0.98	1.00	0.97	1.00	1.00	0.96	1.00	0.97	0.98
12	0.89	0.71	0.85	0.96	1.00	0.97	0.73	0.99	0.92	0.98
13	0.99	0.97	0.95	1.00	1.00	1.00	0.96	0.98	0.99	0.98
14	0.90	0.72	0.85	0.96	1.00	0.98	0.74	0.99	0.92	0.98
15	0.90	0.73	0.82	0.95	1.00	0.98	0.75	0.98	0.92	0.98
16	0.88	0.72	0.80	0.93	1.00	0.96	0.74	0.97	0.89	0.98
17	0.92	0.92	0.96	0.96	1.00	0.98	0.94	0.99	0.95	0.98
18	0.98	0.98	0.99	0.99	1.00	1.00	0.97	1.00	0.99	0.98
19	0.97	0.96	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.98
20	1.00	0.99	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.98
21	1.00	0.94	1.00	1.00	1.00	1.00	0.95	1.00	1.00	0.98
22	1.00	0.99	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.98
23	0.96	0.95	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.97
24	0.92	0.91	1.00	1.00	0.95	0.98	0.94	1.00	1.00	0.95
25	0.96	0.95	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.97
26	0.91	0.90	1.00	1.00	0.94	0.98	0.93	1.00	1.00	0.94
27	0.98	0.98	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.98
28	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99
29	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99

The unequal-rates model is highlighted in grey shading. Figure 4a shows where the nodes are located on the phylogeny.

nature research

Corresponding author(s): Ehab Abouheif

Last updated by author(s): May 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Wide field fluorescence and DIC imaging of embryos was done using Zeiss Axiovision v4.9.1, Confocal miscroscopy was done using Leica SP8. qPCR data was collected by the Genomic Platform facility at the Institute for Research in Immunology and Cancer at the University of Montréal (Quebec, Canada) using a QuantStudio 7 Flex Real-Time PCR System.
Data analysis	DNA alignments, gene trees and BLAST analysis were performed on Geneious (R8). All statistical analyses were performed using Prism GraphPad v7 or Microsoft Excel (2013). Phylogenetic analyses were performed using RevBayes v1.7.1 and Tracer 1.7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequence data that support the findings of this study have been deposited in Genbank with accession code MH801205 and in NCBI Sequence Read Archives with the accession codes [PRJNA625680, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP256614].

All raw qPCR data has been provided as a Source Data File for Extended Data Figure 4q, r called 'Source Data Extended Data Fig. 4'

All raw image data that support the findings of this study are publicly available in figshare with the the following identifiers:
>>>Reference number 78072
https://figshare.com/projects/The_origin_and_elaboration_of_a_major_evolutionary_transition_in_ants/78072
Figure 1: 10.6084/m9.figshare.12133308
Figure 2: 10.6084/m9.figshare.12133311
Figure 3: 10.6084/m9.figshare.12133314
Figure 4: 10.6084/m9.figshare.12133326
Extended data figure 1: 10.6084/m9.figshare.12133296
Extended data figure 2: 10.6084/m9.figshare.12133287
Extended Data figure 3: 10.6084/m9.figshare.12133110
Extended data figure 4: 10.6084/m9.figshare.12133278
Extended data Figure 5: 10.6084/m9.figshare.12130902
Extended data figure 6: 10.6084/m9.figshare.12131022
Extended data figure 7: 10.6084/m9.figshare.12132993
Extended Data figure 8: 10.6084/m9.figshare.12131430

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	(1) For all statistical analyses, our ability to detect a significant difference between controls and experimental treatments at a significance level of alpha = 0.05 suggest that samples sizes used were sufficient. (2) For RNAi experiments for qPCR analysis: for both RNAi treatment and its control YFP RNAi, 40 individual dissected bacteriocytes were pooled into 4 replicates of 10 each. Similarly, 40 individual dissected capsules were pooled into 4 replicates of 10 each, and 40 individual dissected germbands and yolk sac were pooled into 4 replicates of 10 each. (3) For RNAi experiments for phenotype analysis: a minimum of 100 embryos was deemed sufficient for each gene targeted. The final sample size for each experiment was subject to variable mortality during the culturing process. The number of treated embryos (abdA or ubx RNAi) that exhibited phenotypes are reported as a number and percentage of the total number injected. We did the same for YFP RNAi control embryos. (4) For a given gene, in situ hybridization and immunohistochemistry sample size for C. floridanus consisted of at least 30 embryos or ovarioles of similar stages, while for other species that produce much fewer embryos, the sample size consisted of at least 5 embryos of similar stages. 100% of the embryos sampled showed the same expression patterns.
Data exclusions	(1) For all experiments, those embryos that died or were damaged during the experimental run or while handling post-experiment were excluded to ensure all developmental landmarks and tissues are consistently observable.
Replication	(1) For RNAi experiments for phenotype analysis: we performed each RNAi treatment (in parallel with its control YFP RNAi) at least three times independently. (2) For RNAi experiments for qPCR analysis: for both RNAi treatment and its control YFP RNAi, 40 individual dissected bacteriocytes were pooled into 4 technical replicates of 10 each. Similarly, 40 individual dissected capsules were pooled into 4 technical replicates of 10 each, and 40 individual dissected germbands and yolk sac were pooled into 4 technical replicates of 10 each. (3) In situ hybridization and immunohistochemistry experiments for C. floridanus were repeated at least eight times independently. For other species these were repeated at least four times. (4) For all experiments, all of our attempts at replication were successful.
Randomization	(1) To eliminate any colony or day-of-injection related effects in RNAi experiments for phenotype and qPCR analysis, embryos laid by multiple queens were collected, randomized between treatment and control, and injected on the same day. (2) For immunohistochemistry and in situ hybridization, embryos and ovarioles were collected from different colonies and were randomized before staining.
Blinding	(1) For qPCR, the experiments were performed blind at the Genomic Platform facility at the Institute for Research in Immunology and Cancer at the University of Montréal (Quebec, Canada). (2) Two different researchers independently performed the following steps without communicating each step: sample collection from colonies, randomization of embryos between treatments, treatment of samples, replicate maintenance, and data acquisition and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Primary antibodies:
Rabbit anti-Vasa (1:100), Rabbit anti-Tudor (1:100), Rabbit anti-Germ cell-less (1:300), Rabbit anti-Aubergine (1:50), Rabbit anti-Oskar (1:100) (Source: gift from Paul Lasko Lab, Department of Biology, McGill University).
Mouse anti-HSP90 antibody (1:100) (Source: BD bioscience, Clone# 68, Cat # 610418)
Mouse anti-UbdA antibody (1:4) (Source: DSHB, Clone# FP6.87, Cat# UBX/ABD-A FP6.87, RRID:AB_10660834).

Fluorescent secondary antibodies:
Donkey anti-Rabbit polyclonal Alexa fluor-488 (1:500) (Source: AbCam, Cat# ab150073).
Donkey anti-Mouse polyclonal Alexa fluor-488 (1:500) (Source: AbCam, Cat# ab150105).

Alkaline phosphatase secondary antibodies:
anti-DIG-AP (1:500)(Source: Roche, Cat# 11093274910).
Streptavidin-AP (1:500)(Source: Roche, Cat# 11089161001).

Validation

All antibodies used in this study were validated using Drosophila samples and show conserved patterns of expression (see Supplementary Table 1). Furthermore, immunostains correspond directly with in situ hybridization stains of other germline genes, and for Ubx and abdA, the in situ hybridization stains matched the UbdA antibody (which recognizes both Ubx and abdA). Secondary antibodies did not require validation.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

The study did not involve laboratory animals

Wild animals

The study did not involve wild animals

Field-collected samples

(1) All species collected outside of Canada (see Methods) were imported into Canada under the following import permit numbers from the Canadian Food Inspection Agency: Arizona (P-2016-02919, P-2018-00739), New York (P-2016-02922, P-2018-00737), Texas (P-2016-02918, P-2018-00738), Florida (P-2016-02921, P-2018-00809), Thailand (Ants from Asia, P-2019-00011), Germany (Antstore, P-2019-00293).

(2) All colonies were housed in growth chambers at McGill University's Phytotron Facility under Plant Pest Containment Level 1 certification numbers PC-2016-057 and PC-2018-265 from the Canadian Food Inspection Agency.

(3) Colonies were maintained in plastic boxes with glass test tubes filled with water constrained by cotton wool, and were fed a combination of mealworms, crickets, fruit flies and Bhatkar-Whitcomb diet. All colonies were maintained at 25℃, 70% relative humidity and 12 h day:night cycle.

(4) None of the species collected are endangered as determined by their absence on the IUCN Red List of Threatened Species (https://www.iucnredlist.org/, and search Formicidae).

(5) None of the species were collected on protected lands. Therefore, none of the species collected in Canada, USA, Italy, and Thailand required collecting or export permits. For the species collected in South Africa and China, collecting or export permits were obtained by Antstore, and for species collected in Peru, collecting or export permits were obtained by Andrew Suarez (University of Illinois Ubana-Champaign).

(6) All experiments were performed on female ants.

(7) All colonies were collected from the wild as whole colonies or newly-mated queens using standard ant collecting procedures and

	subsequently transported in plastic containers. They are maintained alive indefinitely in growth chambers at McGill University's Phytotron Facility.
Ethics oversight	Ethics oversight was not required because ants do not require IRB or ethics approval

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Value-guided remapping of sensory cortex by lateral orbitofrontal cortex

https://doi.org/10.1038/s41586-020-2704-z

Received: 16 October 2019

Accepted: 22 June 2020

Published online: 3 September 2020

Check for updates

Abhishek Banerjee[1,3]✉, Giuseppe Parente[1,4], Jasper Teutsch[1,3,4], Christopher Lewis[1], Fabian F. Voigt[1,2] & Fritjof Helmchen[1,2]✉

Adaptive behaviour crucially depends on flexible decision-making, which in mammals relies on the frontal cortex, specifically the orbitofrontal cortex (OFC)[1–9]. How OFC encodes decision variables and instructs sensory areas to guide adaptive behaviour are key open questions. Here we developed a reversal learning task for head-fixed mice, monitored the activity of neurons of the lateral OFC using two-photon calcium imaging and investigated how OFC dynamically interacts with primary somatosensory cortex (S1). Mice learned to discriminate 'go' from 'no-go' tactile stimuli[10,11] and adapt their behaviour upon reversal of stimulus–reward contingency ('rule switch'). Imaging individual neurons longitudinally across all behavioural phases revealed a distinct engagement of S1 and lateral OFC, with S1 neural activity reflecting initial task learning, whereas lateral OFC neurons responded saliently and transiently to the rule switch. We identified direct long-range projections from lateral OFC to S1 that can feed this activity back to S1 as value prediction error. This top-down signal updated sensory representations in S1 by functionally remapping responses in a subpopulation of neurons that was sensitive to reward history. Functional remapping crucially depended on top-down feedback as chemogenetic silencing of lateral OFC neurons disrupted reversal learning, as well as plasticity in S1. The dynamic interaction of lateral OFC with sensory cortex thus implements computations critical for value prediction that are history dependent and error based, providing plasticity essential for flexible decision-making.

Animals adapt their behaviour to variable contextual changes in the environment. Central to adaptive behaviour is value-guided decision making, the ability to flexibly associate stimuli with preferred actions on the basis of past rewards or lack of rewards. Deficits in behavioural flexibility characterize brain disorders such as autism and schizophrenia[1]. In mammals, the prefrontal cortex is the locus of value-guided decision-making[2,3], with the OFC implicated in cognitive evaluation of associations between stimuli and outcomes[4–7]. The OFC is a higher-order area with extensive connections to sensory cortices and subcortical structures of the reward system[8,9]. Yet how neurons in OFC respond to changing reward contingencies is poorly understood. Further, whether OFC neurons instruct sensory areas to remap stimulus–outcome associations in support of adaptive behaviour is unclear.

To study flexible decision-making, we used a reversal learning paradigm based on tactile discrimination. We trained mice to perform a 'go/no-go' texture-discrimination task[10] (with P100 and P1200 sandpaper as the go and no-go textures; Methods; Fig. 1a). Once task performance reached expert level (discriminability index $d' > 1.5$), we implemented a 'rule switch' by reversing the stimulus–reward contingency (Fig. 1b). Mice achieved high d' values during initial learning (from 'learning naive', LN, through 'learning expert', LE), showed decreased performance after reversal and finally re-learned the task (from 'reversal naive', RN, through 'reversal expert', RE) (Fig. 1c, Extended Data Fig. 1, $n = 11$ mice). Reversal learning was faster than initial learning, and performance remained stable over weeks (Fig. 1c, Extended Data Fig. 1). Task performance depended on sensory input and was independent of the initial go texture ($n = 2$ mice; Extended Data Fig. 1). During initial learning, mice developed anticipatory whisking and well-timed licking[11]. After the rule switch, overall whisking behaviour was unchanged, but mice transiently reverted to delayed licking before re-learning (Extended Data Fig. 2). We investigated two brain areas implicated in task learning: the barrel cortex in S1, important for tactile discrimination and sensory learning[12], and the lateral OFC (lOFC), which is critical for the assignment of outcome value[8]. To examine the necessity of these areas for task learning, we expressed inhibitory DREADD receptors (hM4Di) in excitatory neurons in either S1 or lOFC (Methods; histological and electrophysiological validation are shown in Extended Data Figs. 3 and 4). Inhibiting S1 neurons during initial training (via daily CNO injections before each behavioural training session during the LN and LE periods) prevented task acquisition (Fig. 1d). Inhibiting neurons in lOFC, but not medial OFC[7], after the rule switch (RN and RE) impaired reversal learning and increased perseverative errors (Fig. 1d–f, Extended Data Fig. 3). Interestingly, mice with lOFC silencing could still learn a new stimulus–outcome association (of a new texture, P600 sandpaper, with a reward; Fig. 1f). Overall, these results indicate a

[1]Laboratory of Neural Circuit Dynamics, Brain Research Institute, University of Zurich, Zurich, Switzerland. [2]Neuroscience Center Zurich, Zurich, Switzerland. [3]Present address: Biosciences Institute, Newcastle University, Newcastle, UK. [4]These authors contributed equally: Giuseppe Parente, Jasper Teutsch. ✉e-mail: abhi.banerjee@newcastle.ac.uk; helmchen@hifo.uzh.ch

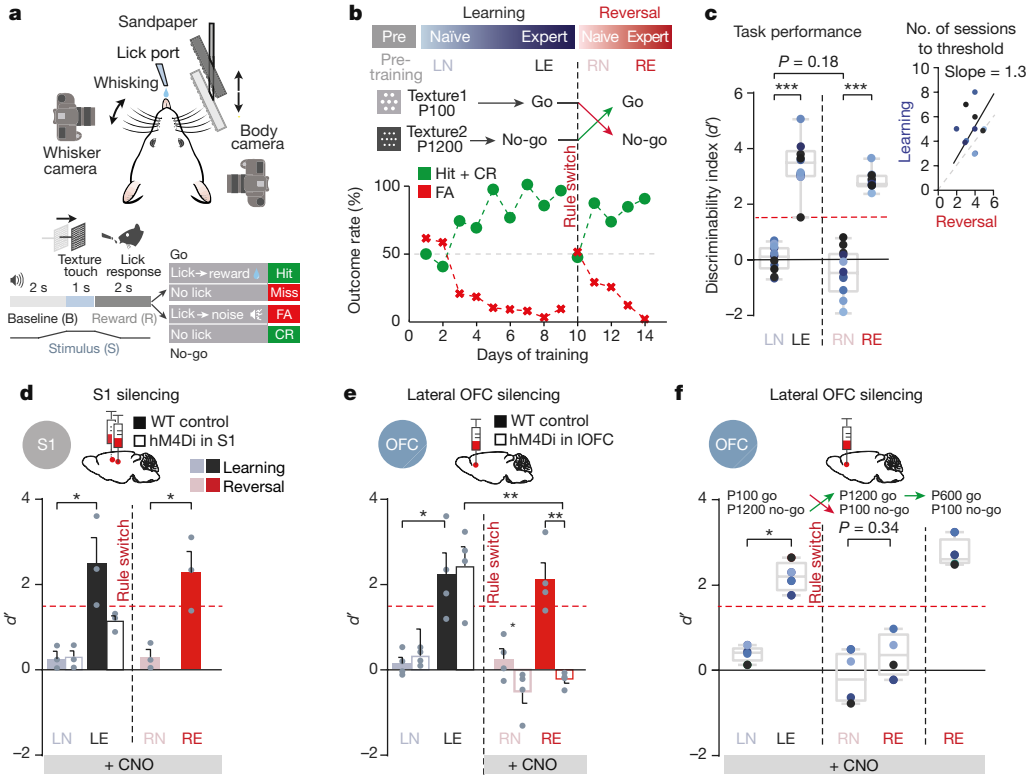


Fig. 1 | Reversal learning is dependent on lateral OFC in a texture-discrimination task. a, Top, schematic of experimental setup. Bottom, trial structure and outcome types (CR, correct rejection; FA, false alarm). **b**, Example of task performance during learning measured as mean correct rate (hit + CR) and false alarm rate. After mice achieved stable high performance, stimulus–reward contingency was reversed (the 'rule switch'). Top, definition of salient task periods (LN, learning naive; LE, learning expert; RN, reversal naive; RE, reversal expert). **c**, Performance (d' values) in the four task periods pooled across 11 mice (dots with different blue shadings). Inset, number of sessions to reach expert level ($d' > 1.5$) for initial versus reversal learning. **d**, In mice expressing inhibitory DREADD (hM4Di) in S1 ($n = 3$), silencing of S1 by systemic application of CNO prevented learning ($d' < 1.5$ in LE; hence mice were not reversed). Control wild-type mice treated with CNO (WT, $n = 3$) learned and re-learned normally. **e**, In mice expressing hM4Di in lOFC, silencing lOFC during RN and RE impaired reversal learning ($n = 4$). **f**, Silencing lOFC throughout all task phases did not affect initial learning but impaired reversal learning ($n = 4$ mice). OFC-silenced mice could still learn a new stimulus–outcome association (a novel P600 go texture). Mean ± s.e.m., $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided Wilcoxon rank-sum test. Box plots show median and 25th and 75th percentiles as box edges, and 5th and 95th percentiles as whiskers.

dissociation of learning and reversal learning mechanisms involving S1 and lOFC, respectively.

To monitor neuronal activity in lOFC and S1 during learning and reversal learning, we performed in vivo two-photon Ca^{2+} imaging in transgenic mice expressing GCaMP6f in excitatory neurons of superficial layer (L) 2/3 of the cortex. We imaged lOFC, located deep in the frontal cortex[13,14], through a gradient-index lens placed in a chronically implanted cannula (Fig. 2a; Extended Data Fig. 5; Methods, $n = 4$ mice). Mice with cannulae implanted showed no impairment in whisking or other behaviours (Extended Data Fig. 5). We observed large calcium transients in lOFC neurons, particularly during the reward-outcome window (Fig. 2a). A longitudinally measured example neuron displayed modest reward-related activity during initial learning (LE), but large and robust responses to unexpected rewards immediately after the rule switch (RN) (Fig. 2b). This activity was transient (RN) and decreased as mice re-learned the task (RE). Averaging across all lOFC neurons revealed the same pattern: a significant increase in the amplitude of reward-related calcium transients after the rule switch (LE→RN; Fig. 2c). These findings are consistent with lOFC encoding deviations from expected outcome value after a rule switch[7]. In agreement with this, the response of lOFC neurons to a third rewarded texture (P600), which was associated with a constant small reward unaffected by reversal, remained unchanged (Extended Data Fig. 6). By contrast, L2/3 neurons in S1, when imaged through a chronic cranial window ($n = 5$ mice), exhibited calcium transients during the stimulus-presentation and reward-outcome windows (Fig. 2d). Responses to the rewarded go texture emerged during learning (LN→LE), decreased after rule switch (RN) and were remapped to the new go texture (RE) (Fig. 2e). Response remapping was significant across S1 L2/3 neurons (Fig. 2f). The same response pattern was found in anatomically identified S1→lOFC projection neurons ($n = 3$ mice; Extended Data Fig. 7). The difference between lOFC and S1 was also evident in the fraction of active neurons in the periods of highest engagement: LE and RN for lOFC but LE and RE for S1 (Fig. 2c, f).

We considered whether neurons selective for rewarded hit trials retain selectivity for the old go texture, or remap to the new go texture after reversal: that is, whether they are more selective for stimulus or for outcome. Longitudinal measurements of lOFC and S1 neurons permitted quantification of their response stability or flexibility upon rule switch. To quantify the response selectivity of active neurons, we defined a hit/CR selectivity index (SI) based on receiver operating characteristic (ROC) analysis[15] (ranging from −1 to 1, permutation test, $P < 0.05$; Methods; Extended Data Fig. 8). We focused on values of SI for the reward-outcome window. The SI per se cannot distinguish between stimulus and outcome selectivity because hit and CR trials differ in both texture type and action outcome; however, comparing SI values before and after a rule switch reveals whether a neuron reverses (stimulus-selective) or maintains (outcome-selective) the sign of its SI during the switch. Figure 3a presents schematically the five major classes of SI changes and their distribution in a 2D plot of values before

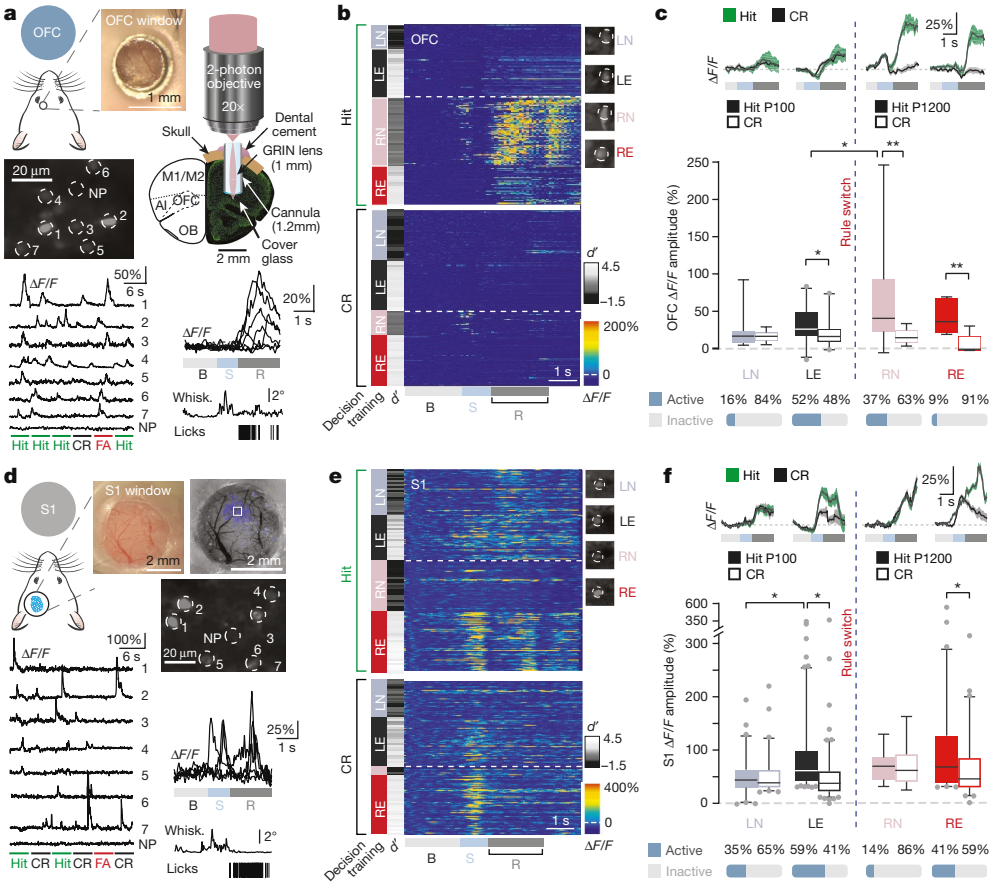


Fig. 2 | In vivo calcium imaging of lOFC and S1 neurons during reversal learning. a, Top, schematic and photograph of cannula window for imaging lOFC. Bottom left, two-photon fluorescence image and GCaMP6f signals (Δ*F*/*F*) during different trial types for example lOFC L2/3 neurons imaged through a GRIN lens. M1/M2, primary and secondary motor cortex; AI, anterior insula; and OB, olfactory bulb. Bottom right, example calcium (Ca^{2+}) transients during hit trials for an individual lOFC neuron with single-trial example of whisking-amplitude and lick events during a hit trial. B, baseline; S, stimulus-presentation window; R, reward-outcome window. **b**, Heat map of single-trial Δ*F*/*F* responses of an example lOFC neuron (sorted by hit and CR; false alarms and misses not shown; performance (*d*′) indicated next to behavioural phases). **c**, Average Ca^{2+} transient amplitude in reward-outcome window for lOFC neurons for hit and CR trials (63 active of 228 recorded neurons in 3 mice; $n = 15$ sessions). Across-trial average Ca^{2+} transients and percentage of active neurons for each phase are above and below graph. **d**, Top, schematic and sample photographs of cranial window above S1. We identified barrel cortex by whisker-evoked intrinsic imaging signals (blue, two-photon imaging area indicated). Middle and bottom left, fluorescence image and GCaMP6f signals (Δ*F*/*F*) for example S1 L2/3 neurons. Bottom right, example Ca^{2+} transients during hit trials for an individual S1 neuron during stimulus and reward-outcome windows; single-trial example of whisking amplitude and lick events below. **e**, Heat map of Δ*F*/*F* transients for an example S1 neuron as in **b**. **f**, Average Ca^{2+} transient amplitude in reward-outcome window for S1 neurons for hit and CR trials (261 active of 539 recorded neurons; 5 mice, $n = 56$ sessions; 11 sessions discarded due to motion artefacts). S1 responses increased in hit trials of both expert phases (LE and RE). Across-trial average Ca^{2+} transients and percentage of active neurons for each phase are above and below graph. Data presented as mean ± s.e.m.; $^*P < 0.05$, $^{**}P < 0.01$, two-sided Wilcoxon rank-sum test. Box plots show median and 25th and 75th percentiles as box edges, 5th and 95th percentiles as whiskers and dots as outliers. Data in **a**, **b**, **d**, **e** are representative of the results shown in **c**, **f**.

versus after the changes; each neuron may have mixed stimulus and outcome selectivity (projections onto the diagonals). To assess both the immediate effect of the rule switch and stable adaptation after re-learning, we classified each neuron into a major class twice (LE→RN and LE→→RE, respectively; Fig. 3a). Among 107 chronically imaged lOFC neurons ($n = 3$ mice), we found a preponderance of outcome-selective neurons that responded strongly to new-hit trials immediately after a rule switch (RN; Fig. 3b, c). Additionally, some lOFC neurons lost or gained selectivity, and this distribution persisted after re-learning (LE→→RE, Fig. 3d; Extended Data Fig. 8). By contrast, S1 neurons were more selective for stimulus than for outcome after reversal (LE→RN, 18% of 218 neurons; $n = 4$ mice; Fig. 3e, f). However, the selectivity of S1 neurons changed markedly during re-learning (LE→→RE), with a large subpopulation functionally remapping to the new, rewarded go texture (Fig. 3g; Extended Data Fig. 8). Moreover, a subpopulation of previously inactive or non-selective neurons acquired outcome selectivity. Similar changes occurred for identified S1→lOFC projection neurons (Extended Data Fig. 7). An analogous analysis of texture-touch-evoked responses during stimulus presentation likewise revealed an overall remapping towards the new go texture (RN→RE, Extended Data Fig. 9). Fitting the data to a generalized linear model[16] further confirmed the link between functional subclasses and behavioural variables, especially reward modulation of outcome-selective neurons (Extended Data Fig. 10; Methods). These results suggest that lOFC neurons exhibit a value-guided response immediately after a rule switch. By contrast, a subpopulation of S1 neurons initially retain the learned association of stimulus with value and functionally remap upon re-learning.

We asked whether delayed S1 remapping is causally dependent on lOFC. To investigate whether OFC→S1 projections existed in mice, we injected retrograde AAV-retro/2-tdTomato into L2/3 of S1. Whole-brain light-sheet microscopy[17] of cleared samples ($n = 2$) revealed dense labelling of S1 projecting OFC neurons, primarily in L2/3 and L5 of the lOFC (Fig. 4a). Chemogenetic silencing of lOFC neurons after the rule switch (RN through RE) impaired remapping of S1 neurons (Fig. 4b;

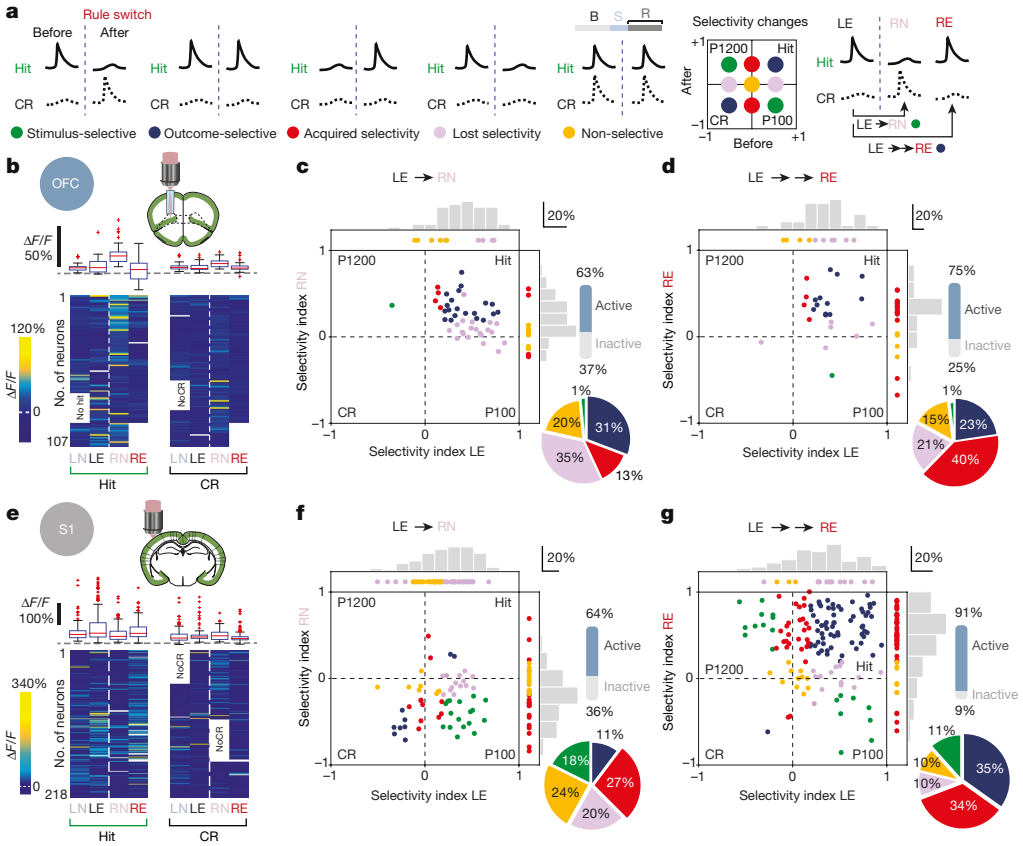


Fig. 3 | Neuronal populations in lOFC and S1 show distinct task-related dynamics. a, Schematic illustrating five major classes of hit/CR selectivity changes upon rule switch and their distribution in a 2D scatter plot of selectivity before and after switch. Right, dual assignment for LE→RN and LE→→RE comparison; selectivity assessed by ROC analysis. **b**, Mean Δ*F*/*F* amplitude in reward-outcome window for lOFC neurons in hit (left) and CR (right) trials, averaged across each salient phase. Bottom, heat maps for 107 longitudinally imaged neurons (20 sessions, 3 mice). Top, box plots of average values pooled across all neurons. **c**, 2D scatter plot and marginal distributions (histograms) comparing hit versus CR selectivity of lOFC neurons in **b** for LE→RN (SI computed in reward-outcome window). Data points above plot, active only in LE; at right, active in RN but not LE; yellow, active with nonsignificant selectivity ($P > 0.05$, permutation test); neurons inactive in both phases not shown (percentage of active neurons is at right). **d**, As in **c** but for LE→→RE. Some lOFC outcome-selective neurons maintained their hit preference, and some previously inactive neurons acquired selectivity for the new hit (51 active of 68 chronically recorded neurons; 16 sessions, 3 mice). **e**, As in **b** but for S1 neurons (218 longitudinally imaged neurons; 28 sessions, 4 mice). **f**, As in **c** but for S1 neurons. Most neurons retained their preference for the previous contingency (90 active of 142 chronically recorded neurons; 20 sessions, 4 mice). **g**, As in **f** but for LE→→RE. Some neurons updated their outcome-selective preference in RE, and some previously inactive neurons acquired new selectivity for the newly rewarded hit trials (198 active of 218 chronically recorded neurons; 28 sessions, 3 mice). Box plots show median and 25th and 75th percentiles as box edges, 5th and 95th percentiles as whiskers and crosses as outliers.

Extended Data Fig. 8; $n = 4$ mice). The effect is best seen in the marginal distributions for the three salient learning periods. Unlike in control mice, a significant fraction of S1 neurons in mice with lOFC silencing preserved their selectivity, failing to remap during re-learning (cumulative distributions, $P < 0.05$, two-sample Kolmogorov–Smirnov test; Fig. 4c). Lateral OFC silencing also prevented RN→RE remapping of responses evoked by texture-touch (Extended Data Fig. 9). We additionally tracked neuronal fate by comparing the assigned classes for LE→RN and LE→→RE transitions. Whereas a fraction of non-selective S1 neurons and of those that had lost selectivity (LE→RN) normally gained selectivity for the new go texture (LE→→RE), such recruitment did not occur in mice with lOFC silencing (Extended Data Fig. 8). These findings further confirm that remapping of S1 crucially depends on top-down input from the OFC.

Finally, we leveraged the sensitivity of lOFC neurons to their history to examine the mechanism by which lOFC influences S1 remapping. Most lOFC neurons that responded to new-hit trials also responded to false alarm immediately after reversal (RN), revealing that lOFC neurons are sensitive to deviations from expected outcome (Fig. 4d, e). We computed a 'reward-history modulation index' (RHMI) for lOFC and S1 neurons by comparing hit trials immediately preceded by a hit or false alarm (Fig. 4f; Methods). Whereas outcome-selective neurons in lOFC exhibited significant response modulation dependent on reward history both before (LE) and after (RN) rule switch, the RHMI was significant in S1 for outcome-selective and acquired-selectivity neurons, but not other classes, after re-learning (RE) (Fig. 4g). History-dependent modulation of S1 neurons was absent in lOFC-silenced mice, indicating that lOFC is critical for the functional reorganization of S1 (Fig. 4g; Extended Data Fig. 10). These findings corroborate the notion that encoding of outcome value by the lOFC is essential to the functional remapping of S1 neurons in support of flexible decision-making.

Adaptive behaviour is shaped by sensory evidence and prediction of the outcome values of future choices. Predictions can shape perception[18], and the OFC estimates the expected value of choices to achieve desirable outcomes, such as increased reward[19]. Our experiments revealed that lOFC neurons have a crucial role in encoding prediction error, which partially resembled classical dopamine responses[20,21]. Critically, lOFC projections to S1 convey this teaching signal, which in turn drives remapping of sensory cortex (Fig. 4h). Tracking both positive and negative outcome values, lOFC neurons may represent ongoing neural estimates of position on a value map[19]. Pharmacogenetic

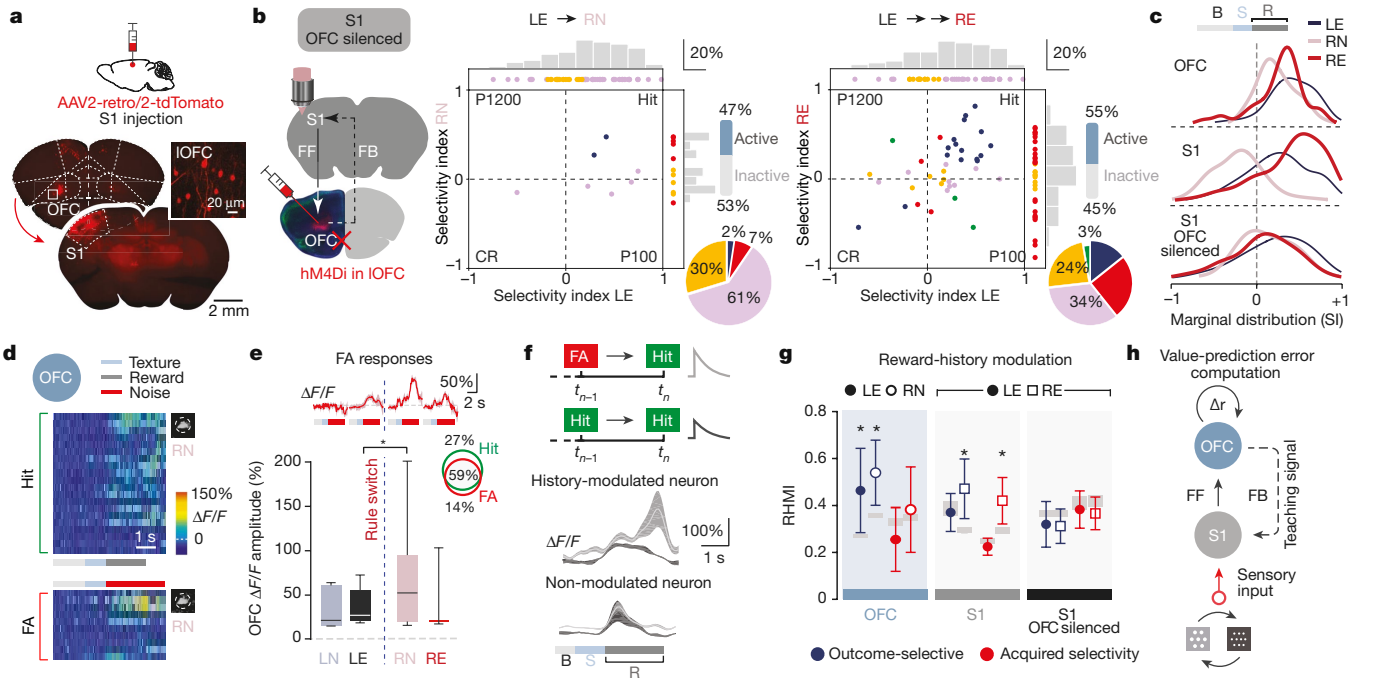


Fig. 4 | Lateral OFC input reconfigures functional responses of S1 neurons. **a**, Retrograde AAV-retro/2-tdTomato injection, CLARITY and whole-brain imaging revealed long-range projections from lOFC to S1 ($n = 2$ mice; inset shows L2/3 lOFC). **b**, Left, schematic of chronic imaging of S1 neurons in lOFC-silenced mice (RN and RE). Middle and right, 2D scatter plots of SI values computed for LE→RN and LE→→RE, respectively; histograms show marginal distributions (85 active neurons of 164 neurons recorded in LE and RN, 24 sessions, 3 mice, 1 session discarded due to motion artefact; 115 neurons of 210 neurons recorded in LE and RE, 25 sessions, 3 mice). **c**, Comparison of SI marginal distributions for LE, RN and RE periods for lOFC and S1 neurons without OFC silencing (Fig. 3c, d, f, g) and S1 neurons in OFC-silenced mice (**b**). **d**, Heat map of single-trial $\Delta F/F$ responses of a lOFC neuron during RN sorted by hit and false alarm (FA) trials. Solid bars indicate periods of texture presentation (light blue), reward (grey) and white noise (red). Data representative of the results shown in **e**. **e**, Average Ca^{2+} transients (top) and mean $\Delta F/F$ amplitudes (bottom) in FA trials for lOFC neurons during four behavioural periods (63 active of 228 neurons in 3 mice). Inset, percentages of active neurons for hit and FA trials, with overlap indicated. **f**, Average hit $\Delta F/F$ responses of two example outcome-selective neurons in S1 exhibiting modulation dependent on trial history, with previous trial rewarded (hit→hit; dark grey trace) or punished (FA→hit; light grey trace). **g**, Reward-history modulation index (RHMI) for outcome-selective neurons (blue) and neurons with acquired selectivity (red) in lOFC, S1 and S1 in lOFC-silenced mice before (LE) and after (RN, RE) rule switch. Data presented as mean ± s.e.m. (*$P < 0.05$; bootstrap-permutation test; s.e.m. of RHMI; permuted indices, grey boxes). **h**, Schematic showing cortico-cortical feedforward (FF) and feedback (FB) interactions for value-prediction error computation in lOFC.

silencing revealed that lOFC is necessary to achieve flexibility, as was previously shown in rodents[22] (although silencing of OFC had mixed effects in non-human primates[6]). Outcome-value signals from lOFC are likely to interact via a rich assortment of projections[23] to integrative cortical areas, such as the retrosplenial cortex[24], and subcortical structures, including the basolateral amygdala[25] and mediodorsal thalamus[26]. Further, we found that a subpopulation of S1 neurons did not function as simple detectors of sensory features, but rather flexibly remapped according to task context and reflect reward history[16,27,28]—characteristics expected in higher-order areas, such as the OFC, but not in primary sensory areas. The cellular and circuit mechanisms enabling this remarkable plasticity remain to be determined but may involve neuromodulators such as serotonin[29] or long-range, layer-specific excitatory and inhibitory interactions[30]. The existence of a signal for reward valence in the primary sensory cortex, and its modulation by higher-order inputs, have important implications for reinforcement learning algorithms[31]. Taken together, this study revealed local and long-range interactions between circuits that are crucial to flexible sensory processing and adaptive decision-making.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2704-z.

1. Fettes, P., Schulze, L. & Downar, J. Cortico-striatal-thalamic loop circuits of the orbitofrontal cortex: promising therapeutic targets in psychiatric illness. *Front. Syst. Neurosci.* **11**, 25 (2017).
2. Miller, E. K. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* **1**, 59–65 (2000).
3. Fuster, J. M. The prefrontal cortex—an update: time is of the essence. *Neuron* **30**, 319–333 (2001).
4. Rolls, E. T. The orbitofrontal cortex and reward. *Cereb. Cortex* **10**, 284–294 (2000).
5. Izquierdo, A. Functional heterogeneity within rat orbitofrontal cortex in reward learning and decision making. *J. Neurosci.* **37**, 10529–10540 (2017).
6. Rudebeck, P. H. & Murray, E. A. The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron* **84**, 1143–1156 (2014).
7. Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T. E. Frontal cortex and reward-guided learning and decision-making. *Neuron* **70**, 1054–1069 (2011).
8. Wallis, J. D. Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.* **30**, 31–56 (2007).
9. Carlén, M. What constitutes the prefrontal cortex? *Science* **358**, 478–482 (2017).
10. Chen, J. L., Carta, S., Soldado-Magraner, J., Schneider, B. L. & Helmchen, F. Behaviour-dependent recruitment of long-range projection neurons in somatosensory cortex. *Nature* **499**, 336–340 (2013).
11. Chen, J. L. et al. Pathway-specific reorganization of projection neurons in somatosensory cortex during learning. *Nat. Neurosci.* **18**, 1101–1108 (2015).
12. Petersen, C. C. H. Sensorimotor processing in the rodent barrel cortex. *Nat. Rev. Neurosci.* **20**, 533–546 (2019).
13. Bissonette, G. B., Schoenbaum, G., Roesch, M. R. & Powell, E. M. Interneurons are necessary for coordinated activity during reversal learning in orbitofrontal cortex. *Biol. Psychiatry* **77**, 454–464 (2015).

14. Jennings, J. H. et al. Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. *Nature* **565**, 645–649 (2019).
15. Pho, G. N., Goard, M. J., Woodson, J., Crawford, B. & Sur, M. Task-dependent representations of stimulus and choice in mouse parietal cortex. *Nat. Commun*. **9**, 2596 (2018).
16. Ramesh, R. N., Burgess, C. R., Sugden, A. U., Gyetvan, M. & Andermann, M. L. Intermingled ensembles in visual association cortex encode stimulus identity or predicted outcome. *Neuron* **100**, 900–915.e9 (2018).
17. Voigt, F. F. et al. The mesoSPIM initiative: open-source light-sheet microscopes for imaging cleared tissue. *Nat. Methods* **16**, 1105–1108 (2019).
18. Bastos, A. M. et al. Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
19. Schoenbaum, G., Roesch, M. R., Stalnaker, T. A. & Takahashi, Y. K. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat. Rev. Neurosci*. **10**, 885–892 (2009).
20. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci*. **23**, 473–500 (2000).
21. Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449–460 (2010).
22. Chudasama, Y. & Robbins, T. W. Dissociable contributions of the orbitofrontal and infralimbic cortex to Pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. *J. Neurosci*. **23**, 8771–8780 (2003).
23. Groman, S. M. et al. Orbitofrontal circuits control multiple reinforcement-learning processes. *Neuron* **103**, 734–746.e3 (2019).
24. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872 (2019).
25. Saez, R. A., Saez, A., Paton, J. J., Lau, B. & Salzman, C. D. Distinct roles for the amygdala and orbitofrontal cortex in representing the relative amount of expected reward. *Neuron* **95**, 70–77.e3 (2017).
26. Rikhye, R. V., Gilra, A. & Halassa, M. M. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci*. **21**, 1753–1763 (2018).
27. Shuler, M. G. & Bear, M. F. Reward timing in the primary visual cortex. *Science* **311**, 1606–1609 (2006).
28. Chéreau, R. et al. Dynamic perceptual feature selectivity in primary somatosensory cortex upon reversal learning. *Nat. Commun*. **11**, 3245 (2020).
29. Bari, A. et al. Serotonin modulates sensitivity to reward and negative feedback in a probabilistic reversal learning task in rats. *Neuropsychopharmacology* **35**, 1290–1301 (2010).
30. Isaacson, J. S. & Scanziani, M. How inhibition shapes cortical activity. *Neuron* **72**, 231–243 (2011).
31. Neftci, E. O. & Averbeck, B. B. Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell*. **1**, 133–143 (2019).

Methods

Animals

All experimental procedures were carried out in accordance with the guidelines of the Federal Veterinary Office of Switzerland and were approved by the Cantonal Veterinary Office in Zurich under license numbers 285/2014 and 234/2018. A total of 30 adult male mice (6–8 weeks of age) were used in this study. For behavioural experiments, we used wild-type (WT) C57BL6/J mice ($n = 16$ mice). For imaging neurons in lOFC and S1, we used *Rasgrf2-2A-dCre*:*CamK2a-tTA*:*TITL-GCaMP6f* triple transgenic mice, which express GCaMP6f in excitatory neocortical layer 2/3 neurons ($n = 14$ mice). For causal pharmacogenetic manipulations, both wild-type (WT) and L2/3-GCaMP6f mice were used ($n = 3$ WT mice and $n = 3$ GCaMP6f mice). To generate triple transgenic mice amenable to two-photon imaging, double transgenic mice carrying CamK2a-tTA (Jackson Laboratories no. 016198[32]) and TITL-GCaMP6f (Jackson Laboratories no. 024103[33]) were crossed with a Rasgrf2-2A-dCre line (Jackson Laboratories no. 022864[34]). The destabilized Cre recombinase expressed under the control of the Rasgrf2-2A promoter was stabilized by trimethoprim (TMP, Sigma T7883) to render it functional. TMP was reconstituted in dimethyl sulfoxide (DMSO, Sigma 34869, 100 mg/ml), freshly prepared before each induction, and administered 2 weeks before surgery. During induction, mice were given a single intraperitoneal injection (150 mg TMP per g body weight diluted in 0.9% saline solution) using a 29-G needle. To specifically label and image from S1→lOFC projection neurons, we injected AAV2.9.hSyn.FLEX.GCaMP6f virus into S1 of WT mice. Mice were grouped with their WT siblings and housed at 24 °C under variable humidity in 12-h reverse dark-light cycle (7:00 to 19:00). At the end of an experiment, the mice were deeply anaesthetized and transcardially perfused or killed with an overdose of pentobarbital (150 mg/kg body weight, i.p.). All efforts were made to minimize suffering. All mice belonged to the C57BL6/J strain.

Reversal learning task

Mice were extensively handled during pre-training sessions to familiarize them with the experimenter and experimental setup. Once they had acclimatized to handling, mice were placed on water restriction and trained on a go/no-go tactile-discrimination task. Mice remained on water restriction for the remainder of the experiment. The behaviour setup has been described previously[10]. During the start of each trial, an auditory cue (2 beeps at 2 kHz, 100-ms duration with 50-ms interval), indicated the approach of one of two possible textures, i.e., sandpapers of grit size P100 (rough texture) or P1200 (smooth texture). The texture was positioned to reach the mouse's whiskers and 'go' or 'no-go' textures were presented pseudorandomly with no more than three consecutive repetitions. The texture stayed in touch with the whiskers for 1 s ('sensation'), after which it moved out of reach. An additional auditory tone (response cue; 4 beeps at 4 kHz, 50-ms duration with a 25-ms interval) signalled the start of a 2-s 'response window' during which the mouse had to lick or refrain from licking the water spout to indicate its choice. A sucrose-water reward was delivered only for licks in response to the 'go' texture and after the response cue ('hit'). Incorrect licks in response to the non-target 'no-go' texture ('false alarms', FA) were punished with a brief period of mild auditory white noise. Reward and punishment were omitted when mice withheld licking for the no-go texture ('correct-rejections', CR) or the go texture ('miss'). The licking detector remained in a fixed and reachable position throughout the entire trial. Animals were motivated to perform the task and typically showed a fraction of 10–15% miss trials during the LN period, which diminished significantly upon learning (LE) and remained low upon rule switch.

Mice proficiently performed the sensory-discrimination task from the learning-naive (LN) through expert phase (LE). Once mice had achieved stable performance of the tactile-discrimination task (reaching $d' \geq 1.5$ for 3 or 4 sessions), the stimulus–response mapping was switched ('rule switch'). Upon rule switch, performance initially dropped to chance level or below. However, after 4–5 days, all mice ($n = 11$ out of 11 mice) learned the new texture-response mapping, increasing performance from reversal naive (RN) through expert phase (RE) as quantified by the increase in d' (training period 4–5 days, 200–300 trials/session/day).

Animal training and performance measurement

We quantified mouse task performance using the discriminability index d-prime (d') rather than the percentage correct to account for motivation and criterion[35]. We set the learning threshold to $d' = 1.5$. d' was calculated for each session as $= Z(\text{hit}/(\text{hit} + \text{miss})) - Z(\text{FA}/(\text{FA} + \text{CR}))$, with $Z(p)$, $p \in [0,1]$, being the inverse of the cumulative Gaussian distribution (FA, number of false alarm trials; CR, number of correct rejection trials). We selected in both training periods, pre- and post-reversal, two relevant salient phases: learning naive and reversal naive (LN and RN, respectively), in which the mice were performing close to or below chance level ($d' \approx 0$, $n = 1$–3 sessions), and learning expert and reversal expert (LE and RE, respectively, $n = 1$–3 sessions), in which the mice were stably performing above $d' = 1.5$. Expert sessions were always selected from the last sessions available immediately before rule switch (LE) or task completion (RE), and this resulted in high performance level ($d' > 2$). For imaging data, only these respective sessions were used.

Whisking and licking measurement

During task performance, whisker kinematics and fine body movement were simultaneously monitored using high-speed cameras. We identified behavioural correlates of task learning by quantifying licking rate and whisking amplitude obtained from lick-sensor measurements and high-speed videography, respectively. The whiskers were illuminated with 940-nm infrared light-emitting diode (LED) light, and movies were acquired during the behaviour at 500 Hz (500 × 500 pixels) using a high-speed CMOS camera (A504k; Basler). Average whisker angle across all imaged whiskers was measured using automated whisker-tracking software. The whisking amplitude (envelope) was calculated as the difference in maximum and minimum whisker angle along a sliding window equal to the imaging frame duration (83 ms). The principal whisker velocity was calculated by applying a bandpass filter to the time vector of the whisking angle and then computing its first derivative. For all trials recorded ($n = 3$ mice), the first and last possible time point for whisker-to-texture contact was quantified manually through visual inspection.

Licking was detected by using a piezoelectric sensor attached to the lick spout, and lick rates were calculated by thresholding this signal and counting the number of events per unit of time. Multiple consecutive threshold crossings that occur in rapid succession can result in a lick rate that exceeds the physical capability of a mouse. We therefore made the reasonable assumption of a peak lick rate of 10 Hz based on manual checks on videography. A low-pass filter was applied to the lick rate time series, which effectively combined multiple events occurring within a 100-ms window into one event. Expert mice showed a decrease in early licks. Although early licks are not exhibited immediately upon rule switch when the behavioural performance is low, lick rates are slightly lower than in expert sessions.

Open-field test

General locomotor activity was measured in an open field (a rectangular arena of 40 × 30 × 20 cm)[36] made from grey Plexiglas that was illuminated from a centred diffuse light source. A single mouse was exposed to the environment for 5 min while being recorded by a video camera placed above the open field and operated by LabVIEW software (National Instruments). Mouse velocity (cm/s) and distance covered (cm) were analysed using EthoVision software.

Horizontal ladder-rung test

A 1-m-long horizontal ladder, consisting of two platforms connected by an irregular pattern of 70 rungs, was used. The distance between rungs varied between 1 and 3 cm. Mice were given time to practice with three trials before they were tested. Three trial sessions per mouse were then recorded using a high-speed camera (Nikon AF Nikkor) at 100 frames per second. Each forepaw placement was analysed, and the quality of the placement was scored using the following scoring system[37]: a perfect paw placement on the rung was scored as 1; partial digit placement, correction and replacement were scored as 0.5; and slip or total miss were scored as 0. The success rate was calculated for each mouse group as

$$\text{Success rate} = (\text{total score}/\text{number of steps}) \times 100 \quad (1)$$

Virus injection

Mice were briefly anaesthetized with isoflurane (2%) in oxygen in an anaesthesia chamber and subsequently transferred to a stereotactic frame (Kopf Instruments). Body temperature was maintained at ~37 °C using a heating blanket with a rectal thermal probe. The eyes of the mouse were covered with vitamin A cream (Bausch & Lomb) during the surgery. The cranium was secured with ear bars, and anaesthesia was maintained during the surgery with 0.8–1.2% isoflurane. After disinfection with Betadine, the skin was opened using a scalpel, an L-shaped incision was made in the skin and the cranial surface was cleaned using absorbent swabs (Sugi; Kettenbach). We identified the lOFC based on stereotactic coordinates from previous studies of 2.6 mm anterior and 1.2 mm lateral from bregma[13]. For S1, injection coordinates were 3.5 mm lateral and 1.3–1.5 mm posterior from bregma. The skull was thinned along a 1-mm line at the rostral edge of S1 using a Dremel drill with occasional cooling with saline. After drilling through the cranium, the dura was punctured using a glass micropipette filled with the virus suspended in mineral oil. Several injections (three or four) were made at neighbouring sites, at a depth of 200–250 µm. A volume of 100–150 nl of virus was injected at a rate of 50 nl/min at each site. After each injection, the pipette was held in place for 5–8 min before retraction to prevent leakage. The skin was sutured using a synthetic, monofilament, non-absorbable suture (Prolene 7.0, Ethicon).

Cranial window and cannula implantation for GRIN lens imaging

To study neural dynamics in the lOFC, a chronically implanted metallic cannula was implanted on top of the lOFC with a glass coverslip at its base. Cannula implantation and cranial window preparation were performed under isoflurane anaesthesia as described above. A circular piece of cranial bone (diameter ~1.5 mm) was removed by drilling on top of the OFC using a Dremel drill. A modified biopsy punch (diameter 1.0 mm; Miltex) was inserted 1.5 mm deep into the cortical tissue for 2 min. The cortical tissue (primary and secondary motor areas) was gently aspirated with a cut using a 27-gauge needle connected to a water jet pump, while being rinsed constantly with Ringer solution. We removed the overlying cortex using aspiration until we reached layer 5 (depth 1.5–1.7 mm) and implanted a stainless-steel cannula (internal diameter 1.0 mm, height 1.5 mm), with its base covered by a cover glass (thickness 0.17 mm), 1.6–1.8 mm below the pial surface. The cannula was secured in place by ultraviolet-light-curable dental acrylic cement (Ivoclar Vivadent). We waited 2–3 weeks after surgery before commencing training. Before each imaging session, a rod-like gradient-index (GRIN) lens (NEM-100-48-00-50-NC, customized needle endomicroscope for two-photon microscopy, ~0.4 pitch, corrected for wavelength λ = 920 nm, diameter = 1.0 mm, length ~4.3 mm; GRINTECH) was inserted through the cannula and neurons were imaged 100–300 µm below. Before each imaging session, the cannula was cleaned with distilled water.

To allow long-term in vivo calcium imaging in S1, a cranial window was implanted over S1 as described previously[10,38]. A metallic head-post for head fixation was glued to the skull, contralateral to the cranial window, using dental acrylic. One week after chronic window implantation, mice were handled daily for 1 week while they became acclimatized to a minimum of 15 min of head fixation.

Brain clearing and light-sheet microscopy

To verify task-relevant projections and connectivity between S1 and lOFC, we injected retrograde AAV-retro/2-shortCAG-tdTomato virus in vivo. Two to three weeks after virus injection, mice were perfused, and the brains subjected to a clearing protocol using CLARITY[39]. After perfusion, the brains were post-fixed for 48 h in a hydrogel solution (1% paraformaldehyde, 4% acrylamide, 0.05% bisacrylamide, 0.25% VA044)[39,40] and then hydrogel polymerization was induced at 37 °C. After the polymerization, the brains were immersed in 40 ml of 8% SDS and kept shaking at room temperature until the tissue had cleared sufficiently (20–40 d depending on the age of the mice). Finally, after 2–4 washes in PBS, the brains were put into a refractive index matching solution (RIMS)[39] for the last clearing step. They were left to equilibrate in 5 ml of RIMS for at least 4 d at room temperature before being imaged.

Cleared brains were imaged using a mesoSPIM light-sheet microscope (www.mesospim.org)[17]. Whole-brain imaging revealed that the lOFC receives direct monosynaptic bottom-up, feed-forward projections from both superficial (L2/3) and mostly deep (L5 and L6) layers of S1. Conversely, a similar injection in mouse S1 (1.3–1.5 mm posterior and 3.5 mm lateral from bregma)[10] revealed superficial cortical L2/3 neurons in mouse S1 receiving direct top-down feedback projections from lOFC.

CNO application

Inhibitory DREADDs (CaMKIIα-hM4D(G_i)-mCherry) were used in the chemogenetic silencing experiments, and neuronal populations of interest were virally transfected with AAV-hM4Di injected unilaterally on the superficial layers (L2/3) of contralateral lOFC and bilaterally to superficial (L2/3) and deeper (L5) layers of S1. Intraperitoneal (i.p.) injection of clozapine *N*-oxide (CNO dihydrochloride, 1–5 mg/kg, Tocris, cat. no. 4936), the ligand that activates hM4Di, silenced the activity of neurons. Clozapine (1–5 mg/kg) was used as control as there are reports that a small proportion of systemically administered CNO is metabolized to clozapine[41].

In vivo electrophysiological recordings

We characterized the pharmacogenetic silencing of lOFC neurons by performing acute in vivo electrophysiology in a subset of the mice that we injected with hM4Di after the completion of the reversal learning protocol. To perform acute recordings, mice were anaesthetized with isoflurane (2% for induction and 0.8% during recording), and their body temperature was maintained stable using a heating pad. A small craniotomy (1-mm diameter) was performed to provide access to the left OFC and the brain was covered with silicon oil. A silver wire was placed in contact with the CSF through a small trepanation (0.5 mm) over the cerebellum to serve as reference electrode. A silicon probe (Atlas Neurotechnologies, 16 linear sites, 100-µm spacing) was implanted through the craniotomy into the left cortical hemisphere, and multi-unit activity (MUA) was recorded from the injection site in the left OFC and surrounding cortex. We waited 30 min to allow the recording to stabilize after implantation of the electrode array. After stabilization, the broadband voltage was amplified and digitally sampled at a rate of 30 kHz using a commercial extracellular recording system (RHD2000, Intan Technologies). The raw voltage traces were filtered offline to separate the MUA (bandpass filter 0.46–6 kHz) using a fourth-order Butterworth filter. Subsequently, the high-pass data were thresholded at 6.5 times the standard deviation across the recording session, and the numbers of spikes in windows of interest were

counted. After a baseline recording period of 30 min, CNO (1–5 mg/kg) was injected (i.p.). During the baseline period (30 min), the average firing rate remained stable, while upon CNO injection the average firing rate in the lOFC decreased steadily over time. Recording electrodes in the lOFC showed a stable, significant decrease in spiking activity 30 min after CNO administration, whereas control electrodes from areas uninfected by the virus did not show any modulation. To combine data across mice, the activity at sites with clear MUA was expressed as a percentage of the baseline value, that is, the average spike rate during the 30-min pre-injection baseline (100%). All multi-units were then combined from the injected or control region, and a *t*-test was performed between the baseline period (−30 to 0 min pre-injection) and the post-injection period (30–60 min after injection).

Intrinsic signal optical imaging

The S1 barrel cortex was identified using intrinsic signal optical imaging in mice anaesthetized with approximately 0.8–1% isoflurane. The cortical surface was illuminated with a 630-nm LED, multiple whiskers were stimulated (2–4 rostro-caudal deflections at 10 Hz) and reflectance images were collected through an objective with a CCD camera (Toshiba TELI CS3960DCL; 12-bit; 3-pixel binning, 4,273,347 binned pixels, 8.6-μm pixel size, 10-Hz frame rate)[42].

Intrinsic signal changes were computed as fractional changes in reflectance (R) relative to the pre-stimulus average (50 frames; expressed as $\Delta R/R$). The centres of the barrel columns corresponding to stimulated whiskers were located by averaging intrinsic signals (15 trials), median-filtering (5-pixel radius) and thresholding to find signal minima. Reference surface vasculature images were obtained using 546-nm LED and matched to images acquired during two-photon imaging.

Two-photon imaging

We used a custom-built two-photon microscope controlled by HelioScan[43], equipped with a Ti:sapphire laser system (approximately 100-fs laser pulses; Mai Tai HP, Newport Spectra Physics), a water-immersion ×16 Olympus objective (340LUMPlanFl/IR, 0.8 numerical aperture, NA) for S1 imaging and an ×20 Leica objective (Leica Plan Apo 0.6 NA) for GRIN-lens-based OFC imaging, galvanometric scan mirrors (model 6210; Cambridge Technology) and a Pockels Cell (Conoptics) for laser intensity modulation.

Based on intrinsic imaging, along with the blood vessel pattern, we targeted specific areas of interest for two-photon imaging of L2/3 neurons in each mouse. We excited GCaMP6f at 940 nm and detected green fluorescence with a photomultiplier tube (Hamamatsu). Images (128 × 64 pixels) were acquired at a 12–15 Hz frame rate, and 10–50 cells per field of view were imaged simultaneously. Single trials of 6–8-s duration were recorded, with 1-s breaks between trials to allow the data to be written to hard disk during intertrial periods.

Calcium imaging analysis

Calcium imaging data were first motion corrected using an online piecewise rigid 2D (planar) method (non-rigid motion correction, NoRMCorre) in MATLAB (Mathworks). Regions of interest (ROI) corresponding to individual neurons were found from both the mean image and the standard deviation image generated from a single-trial time series using ImageJ (US National Institutes of Health). ROI masks were manually selected using an online method (OCIA) in MATLAB, and raw fluorescence time courses ($F(t)$) were then extracted as the (non-weighted) mean pixel value for each ROI. Another fluorescence time course was extracted from the neuropil defined by an ROI selecting a portion of non-somatic tissue in the imaging frame. The neuropil calcium signal never resulted in activity peaks significantly high enough to be classified as an active neuron (see discussion below of criteria for active neurons). The background was subtracted in each channel (bottom first-percentile fluorescence signal across entire time series). The time course of percentage change in fluorescence was calculated by subtracting the baseline fluorescence $F_0(t)$ from $F(t)$, then dividing by $F_0(t)$:

$$\Delta F/F(t) = (F(t) - F_0(t))/F_0(t) \quad (2)$$

$F_0(t)$ was estimated as the mean fluorescence value of the first 1.5 s before tactile stimulus onset. For cells that were not silent in the pre-stimulus window, $F_0(t)$ was instead taken as the eighth percentile of a trailing 1.5-s sliding window.

Alignment of cell masks across days

All analyses for the alignment of cell masks across days were performed manually with the aid of custom MATLAB graphical user interfaces in the OCIA software. To align masks across any pair of daily sessions, we first chose one set for the first day and then imported it onto the single-trial image series of the subsequent days. When displacement occurred, the masks were manually moved to the corresponding neurons. This was done for all pairwise combinations of days. We then manually observed by eye each ROI mask comparing it to both the mean and the standard deviation image of the time series on ImageJ, to confirm the presence of each cell across days. If the z-plane did not match and a cell was not found, it was excluded from further longitudinal analysis.

Criteria for active neurons

To determine if a neuron was active during a time period of interest (stimulus-related and reward-outcome-related responses), we independently tested its evoked response using conservative criteria. For each neuron, we calculated its mean response and its peak value ($\Delta F/F$) during the 0.9-s window after a texture was presented (that is, for the stimulus-presentation window) or during the 1.6-s window after the texture was removed (that is, for the reward-outcome window). A neuron was considered active if all the following criteria were met.

– Its response was significantly ($P < 0.01$, *t*-test) different from the average pre-stimulus baseline response (1.5 s before texture was presented).

– Its mean response (for stimulus-presentation or reward-outcome window) was more than 3 × noise from the baseline. This baseline was calculated by averaging a 35-point sliding window across the trial response and taking the fifth percentile of the mean response distribution. The noise level was taken as the first percentile of the distribution of the standard deviation calculated across the same sliding window.

– Its peak response ($\Delta F/F$) (for stimulus or reward-outcome window) was greater than 25%.

– In the 2D scatter plots of selectivity indices (see below), neurons were considered active if they were active in either of the learning periods considered (for example, LE and RN). In other words, they were considered inactive only if they were inactive in both periods.

Selectivity index

We assessed the selectivity of single-neuron activity for specific trial types using a receiver operating characteristic (ROC) analysis, which quantifies the ability of an ideal observer to discriminate between trial types based on single-trial responses[10,15]. For the purpose of this study, we assessed selectivity for hit versus CR trials. We performed the ROC analysis on the segments of the $\Delta F/F$ transients in the trial period of interest, that is, either in the 2-s-long reward-outcome window or in the 1-s-long stimulus window. Specifically, each trial was assigned a "discrimination variable" score (DV) equal to the dot product similarity of the $\Delta F/F$ segment to the mean $\Delta F/F$ segment for the same trial-type minus the dot-product similarity to the mean for the other trial-type (see also Extended Data Fig. 8). Thus, we computed for hit trials

$$\mathrm{DV}_{\mathrm{hit}} = H_i\left(\overline{C}_{\forall j \neq i} - \overline{C}\right) \quad (3)$$

and for CR trials

$$DV_{CR,i} = C_i\left(\overline{H} - \overline{C}_{\forall j \neq i}\right) \quad (4)$$

where H_i and C_i are the single-trial $\Delta F/F$ segments for the ith hit and CR trial, respectively, and $\overline{H}$ and $\overline{C}$ denote the mean $\Delta F/F$ segments for the respective trial type (excluding the individual trial under consideration). We classified trials as belonging to the go texture or the no-go texture if DV (DV_{hit} or DV_{CR}) was greater than a given criterion. To determine the fraction of trials an ideal observer could correctly classify, we constructed an ROC curve by varying this criterion value across the range of DV. At each criterion value, we plotted the probability that a hit trial exceeded the criterion value against the probability that a CR trial exceeded the criterion value. The area under this ROC curve (AUC) indicates the selectivity for trial type, with an AUC value of 0.5 meaning no selectivity. We defined the 'selectivity index', SI, such that it spanned the range from −1 (CR-preferring neurons) to +1 (hit-preferring neurons) by calculating

$$SI = 2 \times (AUC - 0.5) \quad (5)$$

We tested whether neurons showed trial type selectivity above chance using a permutation test creating 500 permutations with trial type labels randomly shuffled. From these permutations, we created a distribution of indices that could have arisen by chance and considered a neuron's SI value as significant if it fell outside the centre 95% interval of this distribution ($P < 0.05$).

Functional classification of neurons

Neurons that met the activity criteria in at least one of the salient learning periods were classified in different groups according to their hit/CR SI value changes upon rule switch. For each of these neurons we compared the SI value in the pre-reversal period (LE) to the SI value in the two post-reversal periods (RN and RE). This resulted in two classifications for each neuron (for LE→RN comparison and LE→→RE comparison) (Fig. 3a). When two SI values before and after reversal were found to be concordant, that is, of the same sign and significant, a neuron's response was classified as 'outcome selective' for the respective post-reversal phase and the specific trial time window considered (stimulus or reward-outcome). Such a neuron's response amplitude was significantly higher for hit compared to CR trials (or CR compared to hit trials), independent of stimulus identity (in the 2D scatter plots, these neurons are found in the upper right and lower left quadrants). When SI values before and after reversal were discordant, that is, of opposite sign and significant, the neuron's response was classified as 'stimulus-selective' as it switched from hit- to CR-preferring (or CR- to hit-preferring), where the new CR was associated with the same stimulus as the previous hit. In the 2D scatter plot, these neurons are found in the upper left and lower right quadrants. If an active neuron was discriminating above chance during the pre-reversal period LE and lost significant selectivity in the post-reversal period considered (RN or RE), or if it simply became inactive, it was classified as a 'lost selectivity' neuron. Likewise, if an inactive neuron or an active neuron without significant selectivity in the pre-reversal period became active and gained a significant selectivity for the new hit/CR trials, it was included in the 'acquired selectivity' group. Finally, all the active neurons that did not show a significant SI value during either phase (based on permutation tests) were considered 'non-selective'. Each of these neurons was assigned twice to a functional group, in earlier (RN) and later phases of reversal (RE). We tracked the class transition through the course of re-learning using a fate map. For each LE→RN group we showed the fraction of neurons falling into the new LE→→RE classes. Only neurons active during both phases are shown.

Reward-history modulation index

To quantify the effect of previous performance on neural responses, we analysed how response magnitude varied as a result of the outcome of the previous trial (punishment or reward)[16]. We compared the response magnitude of each neuron during a hit trial when the previous trial was a rewarded hit ($R_{hit-hit}$) versus the response magnitude when the previous trial was punished (R_{FA-hit}). To quantify modulation by previous trial history, we created a reward-history modulation index (RHMI) by normalizing the difference between these two history-dependent responses by the mean overall response of all the hit trials:

$$RHMI = \frac{|R_{FA-hit} - R_{hit-hit}|}{R_{hit}} \quad (6)$$

Only cells that were active during a specific phase were included in the RHMI analysis for that respective phase. To check whether a neuron was modulated above chance, a bootstrap permutation test was performed (500 permutations).

Generalized linear model

To estimate the contribution of behavioural and task variables (cue, stimulus onset and offset separated by behavioural response, reward delivery, punishment, licking) to the activity of each neuron, we fit a Poisson generalized linear model (GLM) for each session (MATLAB glmnet package). We first down-sampled deconvolved neural data and all behavioural and task variables to 10 Hz and then smoothed neural activity using a Gaussian filter. Regression functions were created from behavioural and task variables by implementing vectors of Gaussian filters (all filters had a standard deviation of 1 s, overlapping and evenly distributed, 1 Gaussian/3 frames, 100 ms/frame, 144 filters). Each imaging session consisted of 100–120 trials of 6 s each (15 Hz) (training set 75% of each run, testing sets 25%; tenfold cross validated with 11 evenly spaced chunks of trials). We used an elastic net regularization consisting of 99% L2 and 1% L1 methods for each individual neuron. Deviance explained was calculated by comparing the activity predicted by the model to the actual activity calculated using data not used during the fitting procedure. Finally, the contribution of each variable to the neural activity was derived by calculating again the deviance explained using just that variable and normalizing it to the total deviance explained. This is plotted separately for each group of neurons.

Statistical analysis

Statistical analyses are described in the main text and in figure legends. If not stated otherwise, we used non-parametric statistical analyses (two-sided Wilcoxon rank-sum test) or permutation tests to avoid assumptions about the distributions of the data. When assumptions could be made based on previous literature and on small datasets (Fig. 1d, Extended Data Figs. 1c and 5), Student's t-test was used. All statistical analysis was performed using custom-written routines in MATLAB. Quantitative approaches were not used to determine whether the data met the assumptions of the parametric tests. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the finding of this study are available upon reasonable request from the corresponding author.

32. Mayford, M. et al. Control of memory formation through regulated expression of a CaMKII transgene. *Science* **274**, 1678–1683 (1996).
33. Madisen, L. et al. Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron* **85**, 942–958 (2015).
34. Harris, J. A. et al. Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. *Front. Neural Circuits* **8**, 76 (2014).
35. Carandini, M. & Churchland, A. K. Probing perceptual decisions in rodents. *Nat. Neurosci.* **16**, 824–831 (2013).
36. Bailey, K. R. & Crawley, J. N. in *Methods in Behavioral Analysis in Neuroscience* (eds Bailey, K. R., Crawley, J. N. & Buccafusco, J. J.) Ch. 5 (CRC, 2009).
37. Farr, T. D., Liu, L., Colwell, K. L., Whishaw, I. Q. & Metz, G. A. Bilateral alteration in stepping pattern after unilateral motor cortex injury: a new test strategy for analysis of skilled limb movements in neurological mouse models. *J. Neurosci. Methods* **153**, 104–113 (2006).
38. Banerjee, A. et al. Jointly reduced inhibition and excitation underlies circuit-wide changes in cortical processing in Rett syndrome. *Proc. Natl Acad. Sci. USA* **113**, E7287–E7296 (2016).
39. Yang, B. et al. Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell* **158**, 945–958 (2014).
40. Chung, K. et al. Structural and molecular interrogation of intact biological systems. *Nature* **497**, 332–337 (2013).
41. Gomez, J. L. et al. Chemogenetics revealed: DREADD occupancy and activation via converted clozapine. *Science* **357**, 503–507 (2017).
42. Gilad, A., Gallero-Salas, Y., Groos, D. & Helmchen, F. Behavioral strategy determines frontal or posterior location of short-term memory in neocortex. *Neuron* **99**, 814–828.e7 (2018).
43. Langer, D. et al. HelioScan: a software framework for controlling in vivo microscopy setups with high hardware flexibility, functional diversity and extendibility. *J. Neurosci. Methods* **215**, 38–52 (2013).
44. Guo, Z. V. et al. Flow of cortical activity underlying a tactile decision in mice. *Neuron* **81**, 179–194 (2014).
45. Sreenivasan, V. et al. Movement initiation signals in mouse whisker motor cortex. *Neuron* **92**, 1368–1382 (2016).
46. Huber, D. et al. Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* **484**, 473–478 (2012).

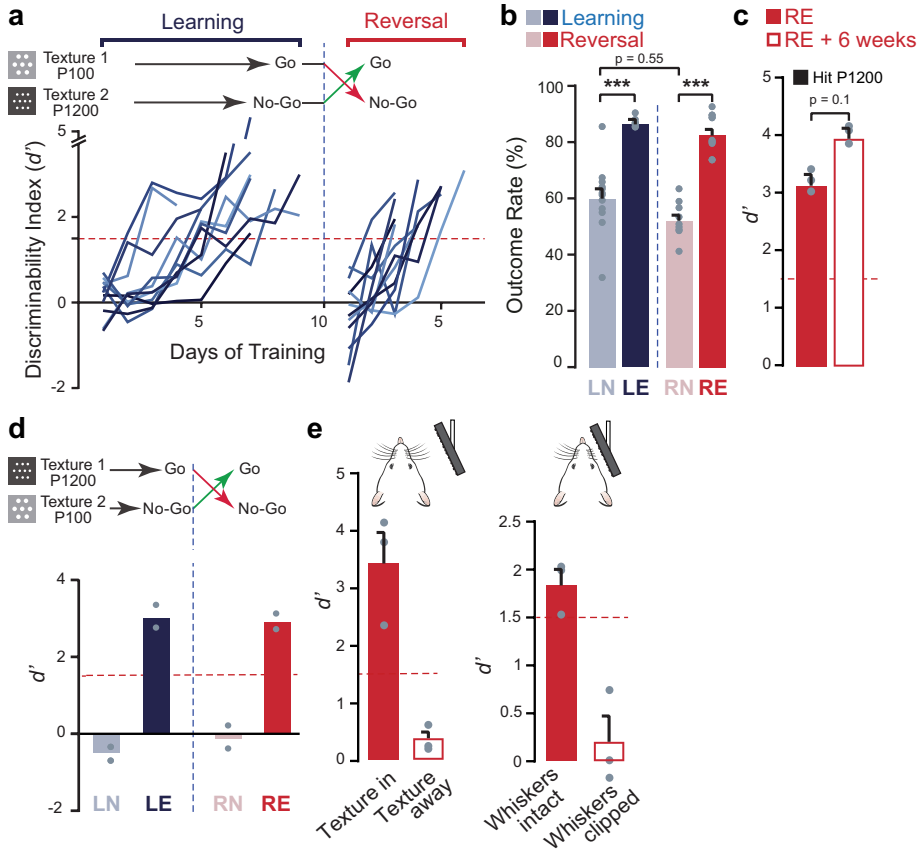
Acknowledgements This work is supported by a H2020 Marie Skłodowska-Curie fellowship (CIRCDYN, grant 709288) and a NARSAD Young Investigator award (grant 24941) from the Brain & Behavior Research Foundation to A.B., and grants from the Swiss National Science Foundation (310030B_170269), Sinergia SNF grant (CRSII5_180316) and European Research Council (ERC Advanced Grant BRAINCOMPATH, 670757) to F.H. We thank B. Grewe for showing us the preparation for GRIN lens imaging; M. E. Schwab for the use of equipment for open-field and ladder-rung test; S. Carta, L. Shumanovski, D. Göckeritz, L. Egolf and C. Rickenbach for various assistance; and W. Senn, F. Lucantonio, M. Goard, M. Pignatelli and B. Scholl for discussions of the manuscript.

Author contributions A.B. conceived the project. A.B. and F.H. designed the study. A.B and J.T. carried out all experiments except multi-unit electrophysiology experiments (performed and analysed by C.L.). A.B and G.P. analysed all data, except whisking kinematics data (analysed by C.L.). F.F.V. developed light-sheet microscope and imaged cleared brains together with A.B.; A.B. and F.H. wrote the manuscript with comments from G.P. and C.L.

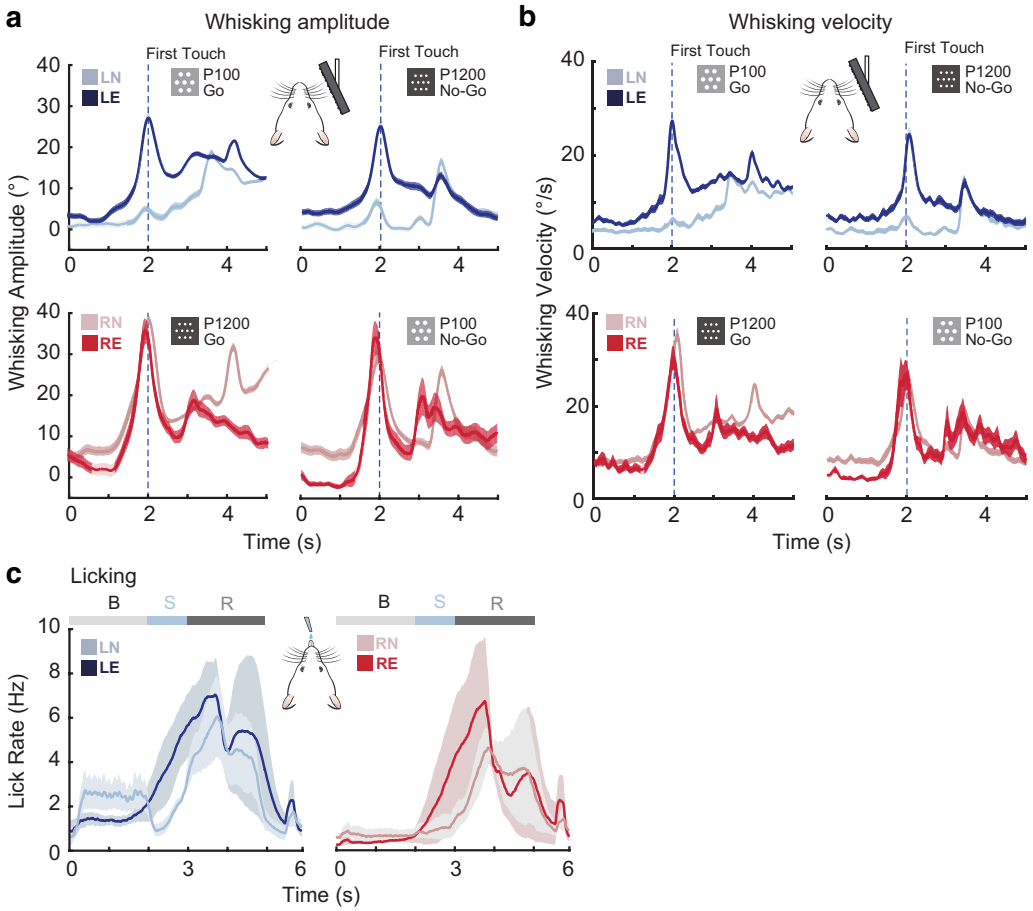
Competing interests The authors declare no competing interests.

Additional information

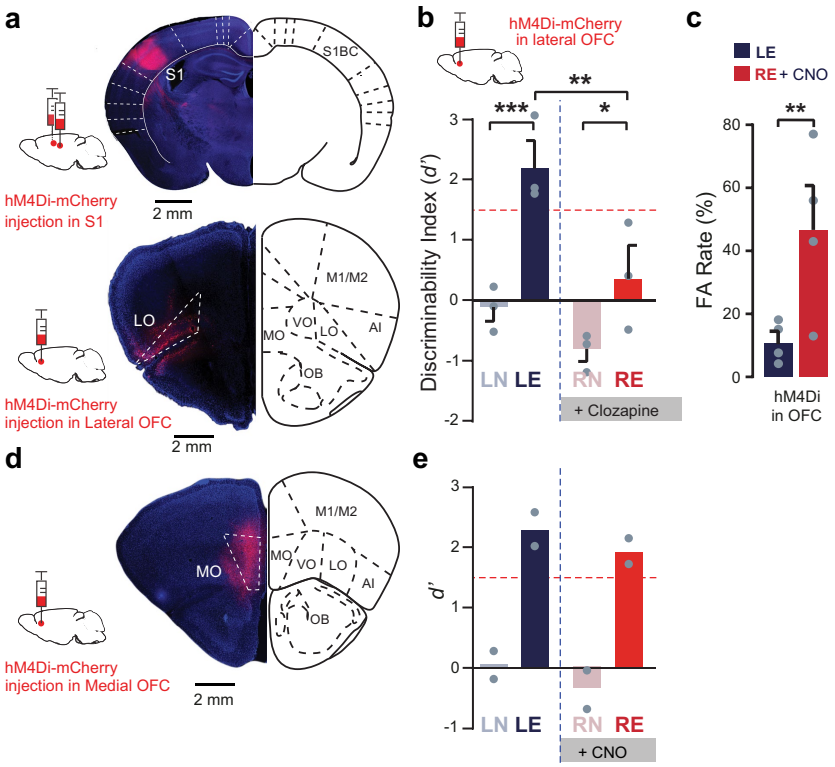
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2704-z.
Correspondence and requests for materials should be addressed to A.B. or F.H.
Peer review information *Nature* thanks Cornelius Schwarz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



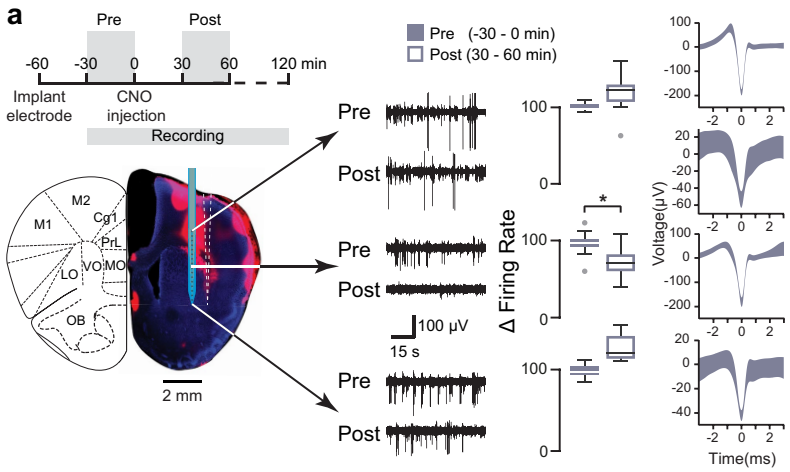
Extended Data Fig. 1 | S1-dependent tactile-discrimination-based reversal learning task. a, Time course of task performance (discriminability index, *d'*) of individual mouse reveals dynamics of learning and reversal learning upon rule switch. Each line in various blue shades represents a single mouse of a total of 11 mice. **b**, Percentage of correct decision '(hit + CR)/all trials' as 'outcome rate' plotted during the four salient behavioural phases of learning (learning naive, LN; learning expert, LE) and reversal (reversal naive, RN; reversal expert, RE) ($n = 11$ mice). **c**, Reversal performance is stable and remains high when mice with reversed reward contingency (P1200 as go texture, RE) were tested 6 weeks later ($n = 2$ mice). **d**, Reversal learning is independent of initial texture training (fine grit size sandpaper P1200 texture as initial go texture; $n = 2$ mice). **e**, Texture discrimination is dependent on sensory input. Left, keeping textures out of reach in expert mice after reversal (RE) impaired their performances ($n = 3$ sessions in 2 mice). Right, clipping whiskers in expert mice similarly resulted in impaired performance (low *d'*) indicating sensory input is essential for the correct execution of the task ($n = 3$ mice, longitudinally studied before and after whisker-clipping). Data presented as mean ± s.e.m., $^{***}P < 0.001$, two-sided Wilcoxon rank-sum test.



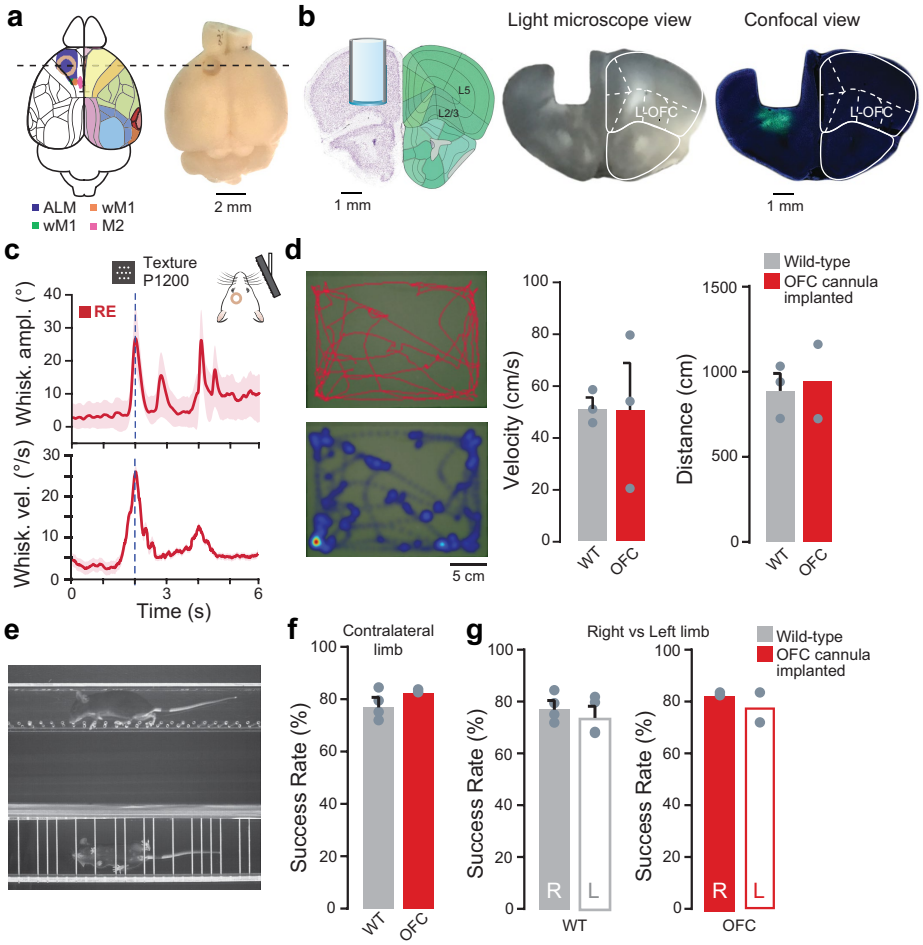
Extended Data Fig. 2 | Whisking and licking behaviour during reversal learning. a, Upper row, time course of envelope whisking amplitude aligned to first-touch during go (left) and no-go trials (right) across two salient periods of initial learning (learning naive, LN; learning expert, LE). Naive (LN) mice showed low-amplitude whisking activity throughout most of the trial. In expert mice (LE), whisking behaviour became time-locked to the arrival of the texture. Lower row, equivalent whisking traces for the periods after rule switch (reversal naive, RN; reversal expert, RE; right). In both RN and RE periods, mice showed stimulus time-locked whisking amplitude ($n = 3$ mice). Note that amplitudes and temporal profiles of the whisking envelope were similar for the smooth P1200 and the rough P100 texture, independent of stimulus-outcome association. **b**, Equivalent analysis as in **a** but for the mean whisking velocity. No significant difference was found in the velocity profile between the two textures in the stimulus-presentation window. **c**, Time-course of average lick rates during go trials across two salient phases of initial learning (left) and reversal learning (right) ($n = 11$ mice). Expert mice (LE and RE) showed both an increase in licking activity during report window (grey) and a decrease of early licks (B, baseline; S, stimulus presentation; R, reward). Data are presented as mean (solid line) ± s.e.m. (shaded area).



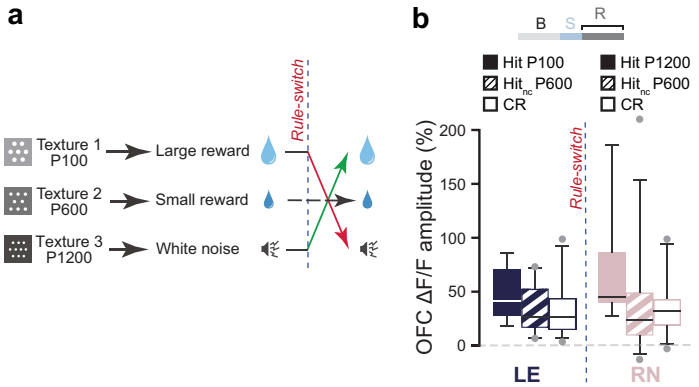
Extended Data Fig. 3 | Immunohistochemical and behavioural validation of pharmacogenetic silencing using hM4Di. a, Neuronal silencing was achieved via viral injection of inhibitory DREADD (AAV-hM4Di-mCherry) into S1 and/or lOFC in mice followed by systemic CNO application. S1 injection (top) was bilateral and lOFC (LO) injection (below) was unilateral and to the ipsilateral side of the barrel field. **b**, Injection of hM4Di in lOFC and systemic administration (i.p.) of clozapine (1–5 mg/kg) after rule switch (RN and RE) selectively impaired reversal learning ($n = 3$ mice). **c**, Injection of hM4Di in lOFC- and CNO-treated mice showed increased perseverative errors (false alarm, FA) in RE compared to LE ($n = 4$ mice). **d**, **e**, Silencing medial OFC (MO) by injecting hM4Di unilaterally in the MO, followed by daily systemic CNO application after rule switch (RN through RE period), did not have any effect on reversal learning. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ two-sided Wilcoxon rank-sum test. Data are presented as mean ± s.e.m.



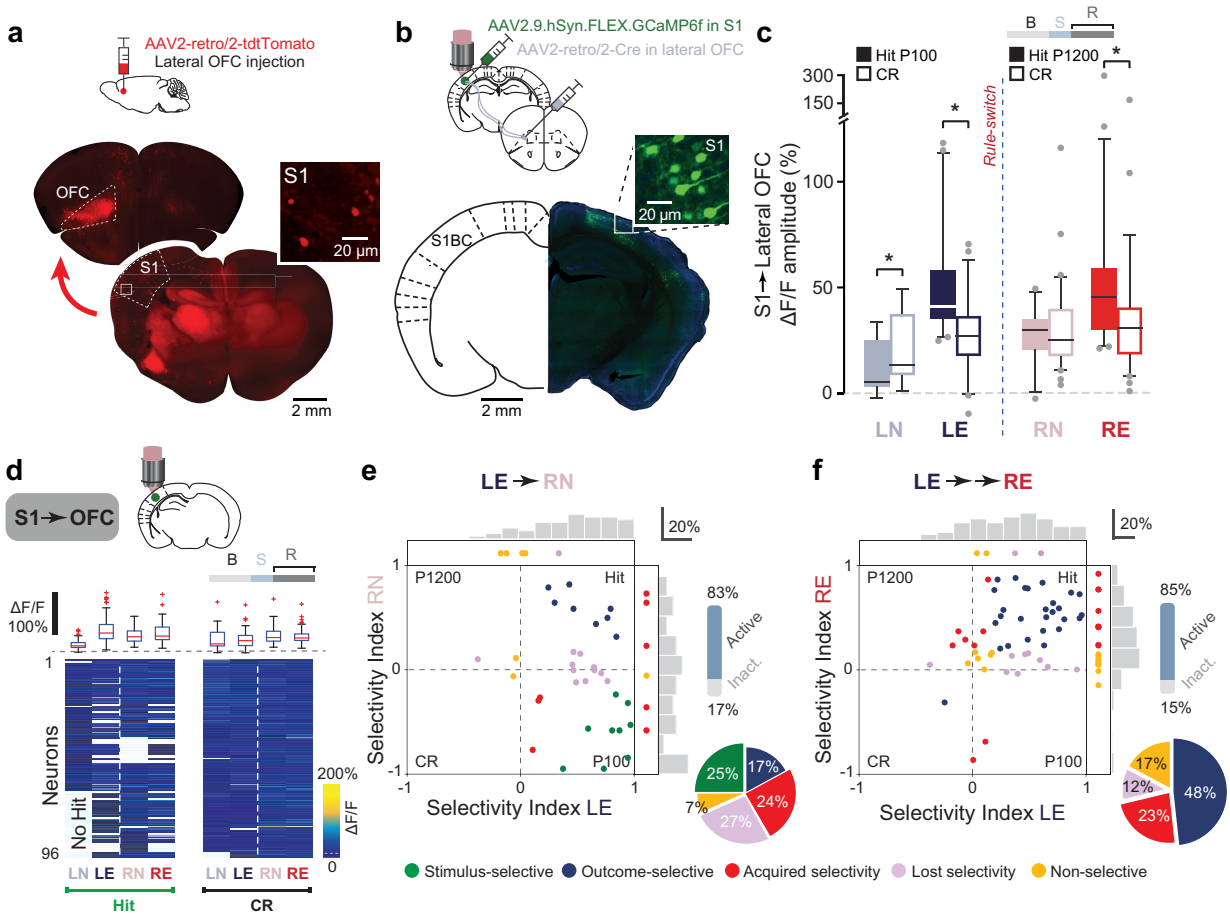
Extended Data Fig. 4 | Electrophysiological validation of lOFC silencing using hM4Di. a, Timeline depicting experimental sequence for validation of lOFC (LO) silencing (top). Schematic of acute electrophysiological recording from frontal cortex (bottom). DAPI stained slice imaged with a confocal microscope showing red fluorescence from DiD to mark the probe location. Example traces from three electrode contacts from one recording session for pre- and post-CNO injection (middle). Box plots showing change in firing rate (% change relative to baseline) for electrode contacts above, in or below lOFC. Plots show median and 25th and 75th percentiles as box edges, and 5th and 95th percentiles as whiskers. To the right, example waveforms from units showing significant modulation by CNO. *$P < 0.05$, *t*-test.



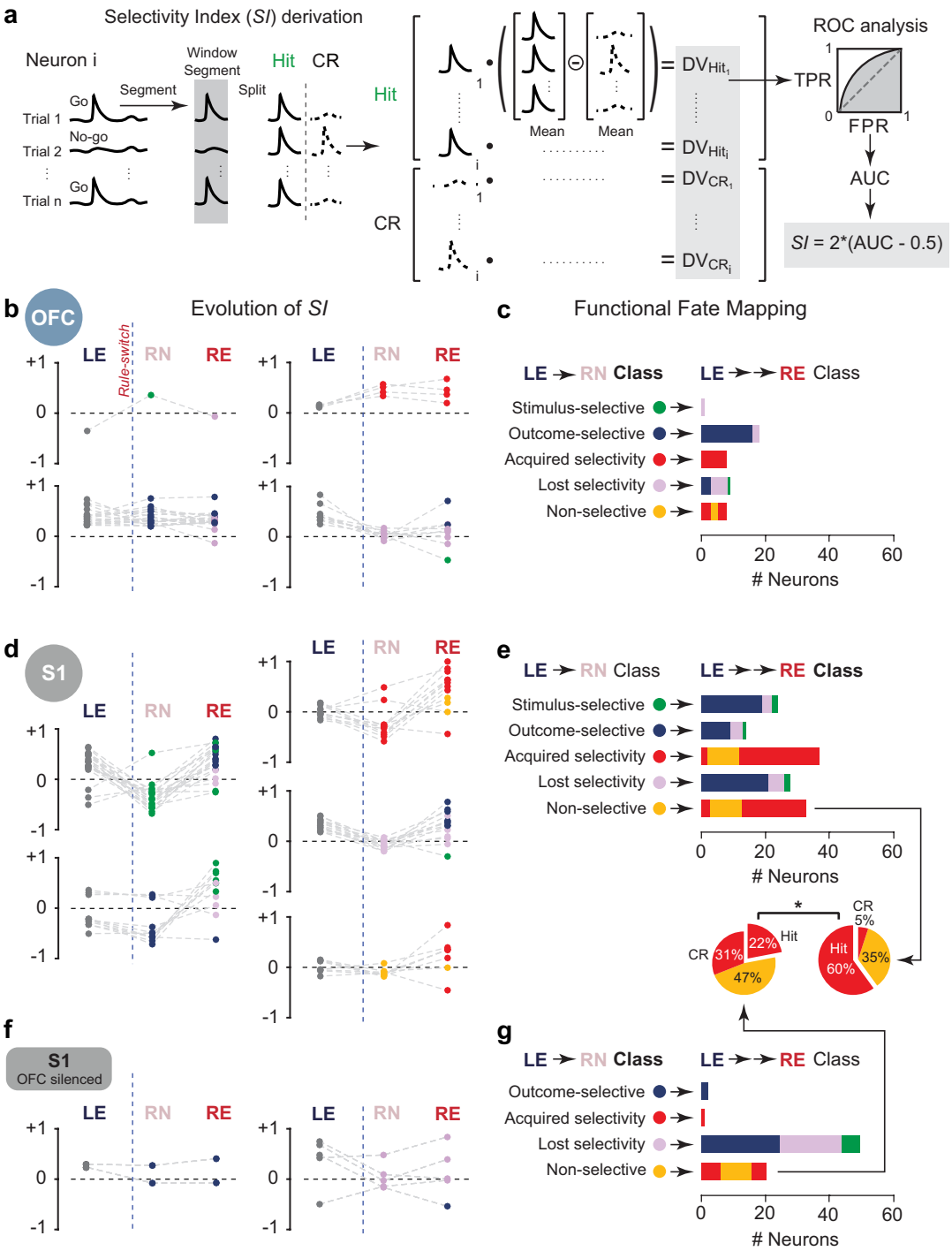
Extended Data Fig. 5 | Unaltered whisking and simple behaviour following OFC cannula implantation. a, A schematic diagram and whole-brain image showing the location of cannula implantation in OFC. Coloured regions on the schematic indicate premotor and motor areas as described in the previous studies[42,44–46] (left hemisphere), or regions according to the Allen institute common coordinate framework (right hemisphere). **b**, A schematic diagram based on the Allen brain atlas and light-microscopic and confocal views shows the GCaMP6f-expressing mice in lOFC (LO) and cannula placement above the virus injection site. **c**, Whisking behaviour is preserved in mice implanted with OFC cannulas. Envelope whisking amplitude (top) and whisking velocity (bottom) in expert mice (RE) centred on the texture approach ($n = 2$ mice). **d**, Open-field test showed normal locomotor function of wild-type non-implanted and OFC cannula-implanted GCaMP6f-expressing mice ($n = 4$ WT and $n = 2$ OFC cannula-implanted mice). Representative picture of locomotor track (top) and heat map (bottom) of an OFC cannula-implanted mouse. Total distance covered (cm) and mean velocity (cm/s) is plotted. Scale bar = 5 cm. **e**, Horizontal ladder-rung test showed normal locomotor function of wild-type (WT, $n = 4$) and OFC cannula-implanted mice ($n = 2$). A representative picture showing paw placement of a mouse on irregular horizontal rung-ladder. **f**, Analysis of paw placement of the limb contralateral to the cannula-implanted side showed no significant difference between WT and OFC cannula-implanted mice. **g**, No differences were seen in paw placement of the limb ipsilateral or contralateral to the cannula-implanted side in OFC cannula-implanted and in control WT mice. Data are presented as mean ± s.e.m.



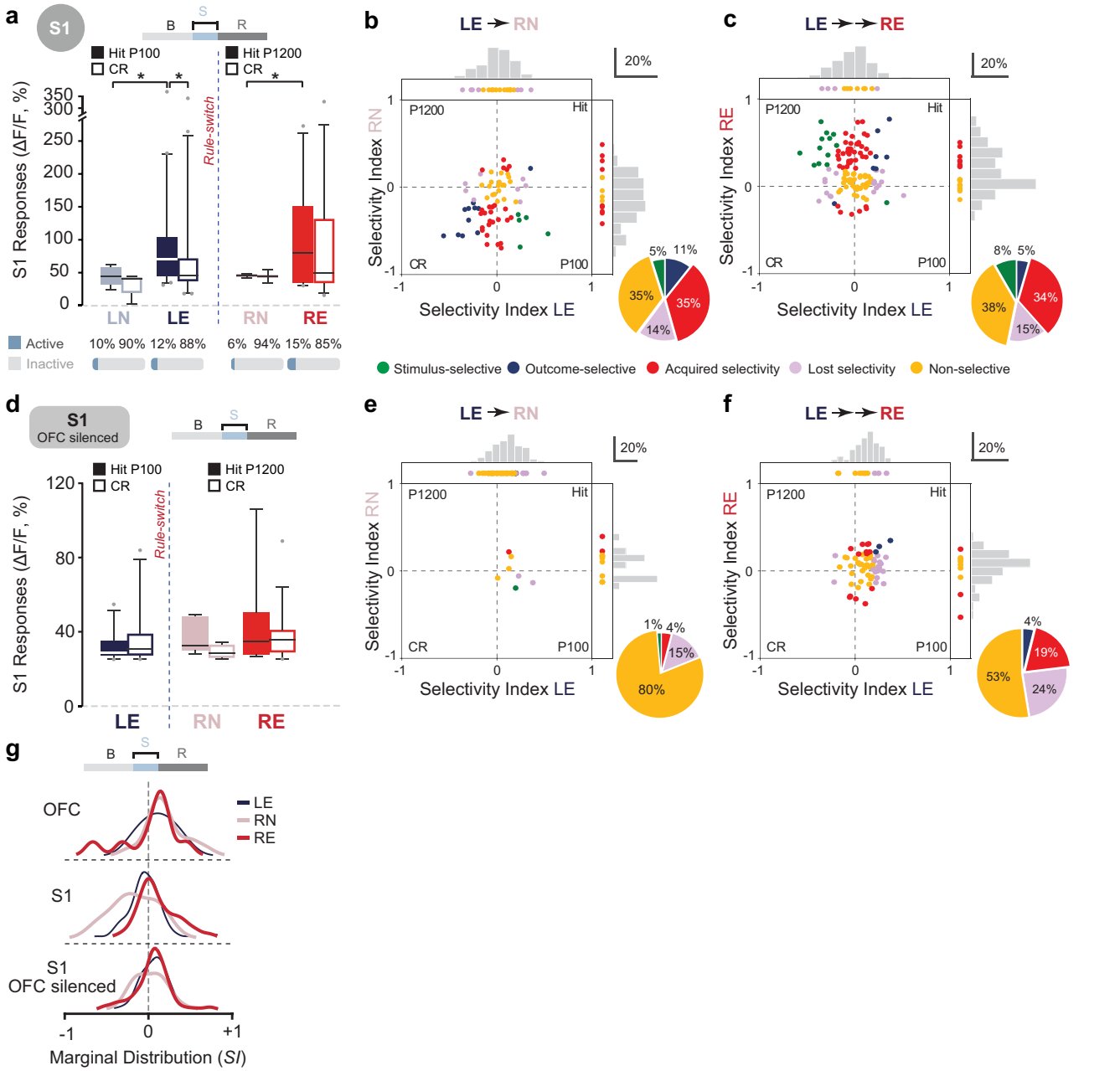
Extended Data Fig. 6 | Re-learning task with neutral context and in vivo Ca^{2+} imaging of lOFC neurons. a, Schematic of the stimulus-outcome associations in a three-textures task with positive (large reward), neutral (small reward), and negative (punishment) context. Same coarse P100 and smooth P1200 sandpapers were used, but an additional intermediate coarseness P600 sandpaper was introduced as go-neutral context (go_{nc}) associated with a small reward, that did not change upon reversal. **b**, Average Ca^{2+} transient amplitude in the reward-outcome window for lOFC neurons for hit, hit_{nc} and CR trials ($n = 63$ active neurons out of 228 neurons recorded in three mice; $n = 15$ sessions) showing increased hit responses upon rule-switch but no significant changes during hit_{nc} trials. Across-trial average Ca^{2+} transients for each behavioural period are shown above. All box plots show median and 25th and 75th percentiles as box edges, and 5th and 95th percentiles as whiskers.



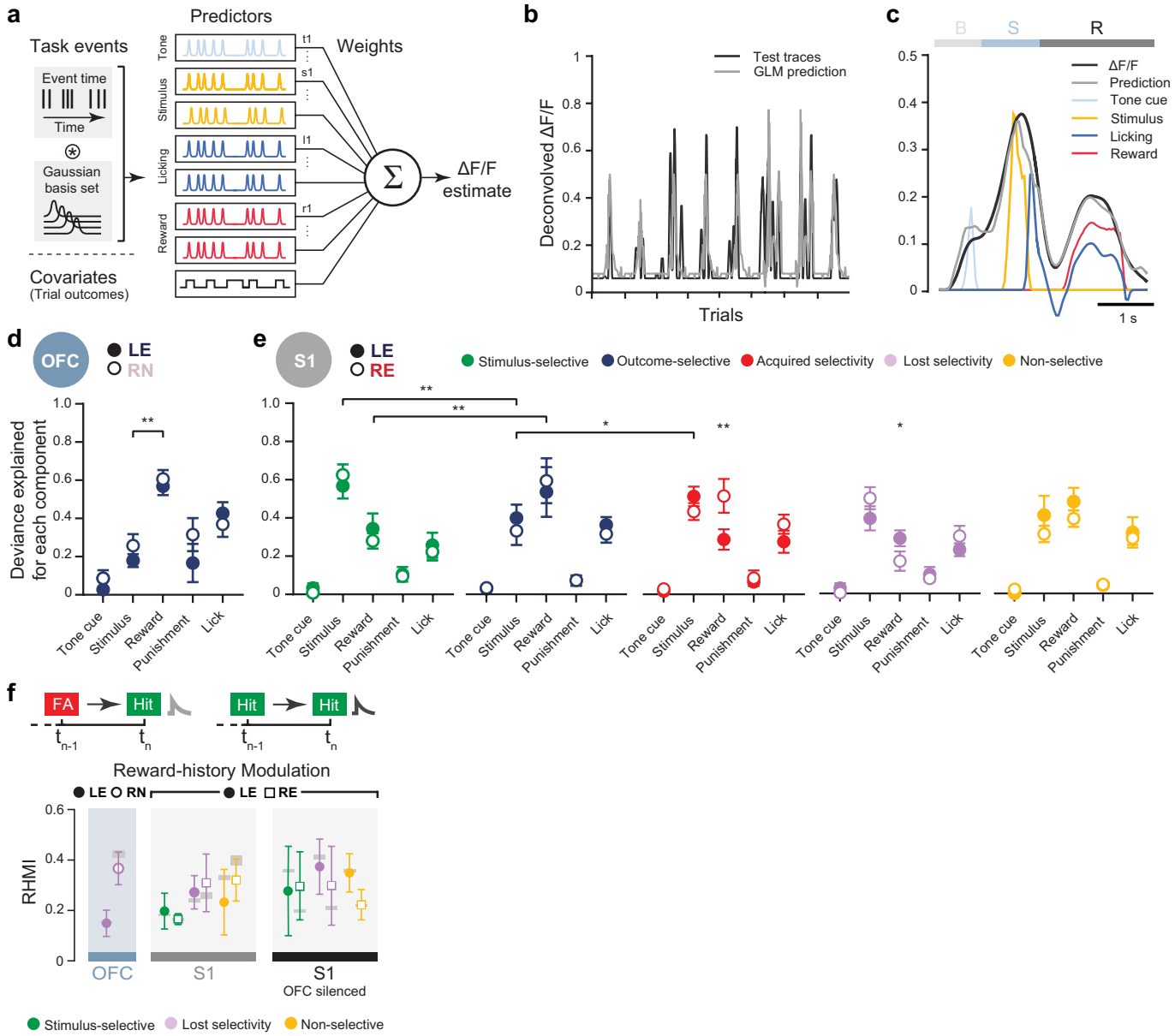
Extended Data Fig. 7 | Task-related functional dynamics in S1→lOFC projecting neurons during reversal learning. a, Retrograde AAV-retro/2-tdTomato injections in vivo in the lOFC followed by clearing the brain using CLARITY and whole-brain light-sheet microscopy revealed feed-forward S1→OFC projections from both deeper (L5 and 6) and superficial (L2/3) layers of S1 ($n = 2$ mice). Labelling is weaker on the contralateral side of the injection site. **b**, S1→lOFC projecting neurons were labelled with GCaMP6f using a dual-viral strategy with retrograde AAV2-retro/2-Cre injected in lOFC and Cre-dependent AAV-DIO-GCaMP6f in S1. Inset, L2/3 neurons in S1 labelled with such strategy. **c**, Average Ca^{2+} transient amplitude in the reward-outcome window shows a significant increase in response amplitude during expert phases of training (LE and RE) ($n = 96$ active neurons over $n = 135$ recorded neurons in 2 mice, $n = 5$ sessions/phase). **d**, Top, S1→lOFC projecting neurons were labelled using a dual-viral strategy with retrograde AAV2-retro/2-Cre injected in lOFC and Cre-dependent AAV-DIO-GCaMP6f in S1. Bottom, peak reward-related responses of S1→lOFC projection neurons averaged across hit (left) and CR (right) trials, longitudinally measured across four salient periods ($n = 96$ neurons from $n = 2$ mice, $n = 5$ sessions/phase). Box plots (median, red line; 25th and 75th percentiles, box edges; most extreme non-outliers, whiskers; outliers, red crosses; zero, dashed grey line) are also shown (inset). **e**, Scatter plot and histogram comparing selectivity index (SI) of S1→lOFC projecting neurons during learning expert (LE) and reversal naive (RN) phase ($n = 39$ active neurons over $n = 46$ neurons from $n = 2$ mice, $n = 5$ sessions/phase). **f**, Scatter plot and histogram comparing SI of S1→lOFC projecting neurons during LE and reversal expert (RE) phases ($n = 61$ active neurons over $n = 73$ from $n = 2$ mice, $n = 5$ sessions/phase). All box plots show median and 25th and 75th percentiles as box edges, and 5th and 95th percentiles as whiskers. Data presented as mean ± s.e.m., $^{*}P < 0.05$, $^{**}P < 0.01$ two-sided Wilcoxon rank-sum test.



Extended Data Fig. 8 | Tracking neuronal responses during early and late phases of reversal learning. a, A schematic view of the step-by-step derivation of the selectivity index (SI) from the ROC curves. **b**, Selectivity indices of longitudinally tracked lOFC neurons across the salient task-periods of LE, RN, and RE. Marker colours for RN and RE indicate the assigned classes for the LE→RN and LE→→RE comparisons, respectively. Plots are shown separately for each LE→RN class. **c**, Fate mapping of longitudinally tracked lOFC neurons. For each LE→RN assigned class, the distribution of these neurons across classes for the LE→→RE comparison is shown as coloured bar on the right. **d**, Same as in **b** but for S1 neurons. **e**, Same as in **c** but for S1 neurons. **f**, Same as in **b** but for S1 neurons in lOFC-silenced mice. **g**, Same as in **c** but for S1 neurons in lOFC-silenced mice. Inset in **e**, the fate distributions of the non-selective neurons in LE→RN show a significantly smaller fraction of neurons that acquire selectivity for the newly rewarded go texture in the RE phase in S1 neurons when lOFC was silenced in mice (22% versus 60%, one-tailed χ^2 test). Note that the fate mapping plots include additional neurons compared to **b**, **d** and **f** as these were not assigned an SI value in each phase but were still classified.



Extended Data Fig. 9 | Texture-touch-related dynamics in S1 neurons during reversal learning. a, Average Ca^{2+} transient amplitude ($\Delta F/F$) in the stimulus-presentation window for S1 neurons ($n = 142$ neurons in $n = 3$ mice, $n = 2$ sessions/phase). **b**, Scatter plot and histogram comparing texture touch-related selectivity index (SI) for the stimulus-presentation window for S1 neurons during learning expert (LE) and reversal naive (RN) phase ($n = 218$ from $n = 3$ mice, $n = 28$ sessions). **c**, Scatter plot and histogram comparing SI of S1 neurons during LE and reversal expert (RE) phase ($n = 218$ neurons from $n = 3$ mice, $n = 28$ sessions). **d**, Average Ca^{2+} transient amplitude ($\Delta F/F$) in the stimulus-presentation window for S1 neurons in lOFC silenced mice ($n = 87$ neurons in $n = 2$ mice, $n = 2$ sessions/phase). **e**, Scatter plot and histogram comparing texture touch-related SI of S1 neurons during LE and RN phase in lOFC-silenced mice ($n = 165$ neurons, $n = 25$ sessions per phase). **f**, Scatter plot and histogram comparing touch-related SI of S1 neurons in lOFC silenced mice during LE and RE phase ($n = 210$ neurons in $n = 3$ mice, $n = 28$ sessions). **g**, Comparison of SI marginal distributions for the three salient periods LE, RN, and RE for lOFC neurons (2D scatter plots not shown), S1 neurons (**c**, **d**) and S1 neurons in lOFC-silenced mice (**e**, **f**). All box plots show median and 25th and 75th percentiles as box edges, and 5th and 95th percentiles as whiskers. *$P < 0.05$, two-sided Wilcoxon rank-sum test.



Extended Data Fig. 10 | Differential modulation of task variable-relevant events in neuronal responses. a, Schematic diagram of a generalized linear model (GLM, Poisson regression) to predict neural activity from behavioural task variables. Each event was expanded into a series of evenly spaced gaussian filters. **b**, GLM predicting deconvolved neural activity of an example S1 outcome-selective neuron from task variables. **c**, Separate components contributing to the average response of this neuron reveal major sensory modulation together with reward-evoked activity. B, baseline; T, texture touch; R, reward. **d**, To quantify each task variable contribution, the relative fraction of deviance explained is calculated and normalized by the total deviance explained for each neuron both before and after reversal. The reward component in lOFC outcome-selective neurons is significantly greater than the touch related component. **e**, Fraction of deviance explained for each component in separate subsets of S1 neurons reveal distinct modulations for specific task-related events. Notably, responses of outcome selective S1 neuronal responses are mostly explained by reward component. Licking activity seems to modulate S1 neural responses less than reward in each subset. Neurons analysed using GLM are same neurons from Fig. 3. Data are presented as mean ± s.e.m., $^{*}P < 0.05$, $^{**}P < 0.01$, two-sided Wilcoxon rank-sum test. **f**, Reward-history modulation index (RHMI) for functional subclasses of lOFC neurons and S1 neurons in OFC intact control mice and lOFC-silenced mice (neurons analysed are from Fig. 4b; ns = $P > 0.05$; bootstrap-permutation test; s.e.m. of RHMI with permutated indices as grey bars).

nature research

Corresponding author(s): Abhishek Banerjee

Last updated by author(s): May 22, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavior set up was programmed as a custom software written in LabVIEW 2011 (64 bit) using commercial library of functions from Phidgets and Zaber libraries. Two-photon calcium data was collected using HelioScan software. Whisker tracking was performed with Noldus EthoVision XT software.
Data analysis	Matlab R2015b (Mathworks), Graph Pad Prism v8 (GraphPad Software), ImageJ Fiji. Behavioral data (animal performance, licking rate, whisking angle) was analysed in Matlab using custom codes that are described in detail in the Methods on pg.15-16. Imaging data was motion corrected using NoRMCorre (Flatiron Institute) available on GitHub. Calcium data was extracted and preprocessed using a custom Toolbox in Matlab (OCIA, HelmchenLabSoftware) available on GithHub. Calcium data was deconvolved using a custom algorithm in Matlab (constrained-foopsi, epnev) available on GitHub. The GLM analysis was performed using the Glmnet Matlab package available on web.stanford.edu. The various indices (Selectivity Index, Reward history modulation Index) were computed using Matlab custom codes as described in detail in Methods. The script library for data analysis will be available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data will be available from the corresponding author upon request. All analysis codes will be made available before publication via Github.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen in accordance with the standard number of animals in comparable studies in the field. Three to six mice per group were used in each imaging experiment. A larger number of mice were used in purely behavioral experiments. The exact number of animals are given in the main text and also in figure legends.
Data exclusions	Mice that lost their hooks during the behavioural training period and longitudinal imaging were excluded from the analyses. Trials or parts of trials with uncorrected large XY-motion or large Z-motion were also excluded. Neurons were excluded from longitudinal analyses when it was not possible to track them in consecutive training phases. The criteria for exclusion of specific neurons in longitudinal recording experiments is given in the Methods section. Exclusion criteria was not pre-established.
Replication	All experiments were reproduced multiple times in multiple animals. All attempts of replication was successful.
Randomization	No randomization was performed as this is not strictly relevant to this study.
Blinding	CNO injection experiments were performed blindly together with WT control.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Adult transgenic male mice (1.5-3 months old) were used in this study. These mice were triple transgenic Rasgrf2-2AdCre; CamK2a-tTA;TITL-GCaMP6f animals. Details are in the Methods section.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	Methods were carried out according to the guidelines of the Veterinary Office of Switzerland and following approval by the Cantonal Veterinary Office in Zurich.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Structure of the C9orf72 ARF GAP complex that is haploinsufficient in ALS and FTD

https://doi.org/10.1038/s41586-020-2633-x

Received: 18 December 2019

Accepted: 20 May 2020

Published online: 26 August 2020

Check for updates

Ming-Yuan Su[1,2], Simon A. Fromm[1,2], Roberto Zoncu[1,2] & James H. Hurley[1,2,3]✉

Mutation of *C9orf72* is the most prevalent defect associated with amyotrophic lateral sclerosis and frontotemporal degeneration[1]. Together with hexanucleotide-repeat expansion[2,3], haploinsufficiency of *C9orf72* contributes to neuronal dysfunction[4–6]. Here we determine the structure of the C9orf72–SMCR8–WDR41 complex by cryo-electron microscopy. C9orf72 and SMCR8 both contain longin and DENN (differentially expressed in normal and neoplastic cells) domains[7], and WDR41 is a β-propeller protein that binds to SMCR8 such that the whole structure resembles an eye slip hook. Contacts between WDR41 and the DENN domain of SMCR8 drive the lysosomal localization of the complex in conditions of amino acid starvation. The structure suggested that C9orf72–SMCR8 is a GTPase-activating protein (GAP), and we found that C9orf72–SMCR8–WDR41 acts as a GAP for the ARF family of small GTPases. These data shed light on the function of C9orf72 in normal physiology, and in amyotrophic lateral sclerosis and frontotemporal degeneration.

Expansion of hexanucleotide GGGGCC repeats in the first intron of *C9orf72* is the most prevalent genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal degeneration (FTD), and accounts for approximately 40% of familial ALS, 5% of sporadic ALS and 10–50% of FTD[1]. Two hypotheses—which are not mutually exclusive—could explain how the mutation leads to a progressive loss of neurons. The toxic gain-of-function hypothesis suggests that toxic molecules, including RNA and dipeptide-repeat aggregates, disrupt neural function and lead to their destruction. The loss-of-function hypothesis is based on the observation of a reduction in *C9orf72* mRNA and protein levels in patients. The endogenous function of C9orf72 is essential for microglia[4] and for the normal dynamics of axonal actin in motor neurons[5], and restoring normal expression of C9orf72 rescues function in *C9orf72*-mutant model neurons[6].

C9orf72 contains longin and DENN domains[7] (Fig. 1a), and exists as a stable complex with another protein that contains these domains, Smith–Magenis syndrome chromosome region, candidate 8 (SMCR8), as well as the WD repeat-containing protein 41 (WDR41)[8–13] (Fig. 1a). WDR41 targets C9orf72–SMCR8 to lysosomes[14] via an interaction with the transporter PQ loop repeat-containing 2 (PQLC2)[15]. Previously proposed functions of C9orf72–SMCR8 include the regulation of RAB-positive endosomes[16], regulation of RAB8A and RAB39B in membrane transport[8,12], regulation of the ULK1 complex in autophagy[9,12,13,17] and regulation of mTORC1 at lysosomes[10,11,18]. Thus far it has been difficult to deconvolute which of these roles are direct and which are indirect. To gain more insight, we reconstituted and purified the complex, determined its structure and assessed its function as a purified complex.

We expressed and purified full-length human C9orf72–SMCR8 and C9orf72–SMCR8–WDR41 (Extended Data Fig. 1a–c). We determined the structure of C9orf72–SMCR8–WDR41 at a resolution of 3.8 Å by cryo-electron microscopy (cryo-EM) (Fig. 1b, c, Extended Data Figs. 2, 3, Extended Data Table 1). We were able to visualize the ordered, approximately 120-kDa portion of the complex, which corresponds to about 60% of the total mass of the complex. Portions of the density—notably, in the DENN domains of both C9orf72 and SMCR8—were very well-resolved, such that side-chain density was clear. Other regions (particularly the longin domains of C9orf72 and SMCR8, and the portion of WDR41 most distal to SMCR8) were less well-resolved, and were not clear enough for side-chain placement. The structure has the shape of an eye slip hook with a long dimension of about 140 Å (Fig. 1c). The ring of the hook was straightforward to assign to WDR41 by its appearance as an eight-bladed β-propeller. The remainder of the density showed evidence of two longin domains at the tip of the hook, with the bulk of the hook made up of two DENN domains. The DENN domain of SMCR8 is in direct contact with WDR41, whereas C9orf72 has no direct contact with WDR41. We assigned the hook-tip portion of the longin domain of SMCR8 to residues I165–A219, which were predicted to comprise a long helical extension unique to this domain. The longin and DENN domains of SMCR8 are near each other but not in direct contact, and are connected by a helical linker that consists of residues K320–V383. Both the longin and the DENN domain of C9orf72 are positioned between the longin and DENN domains of SMCR8. This linear arrangement of domains gives the overall complex an elongated shape.

To map the interactions of WDR41 and to facilitate the interpretation of the less well-resolved portions of the cryo-EM structure, we subjected C9orf72–SMCR8 and C9orf72–SMCR8–WDR41 to hydrogen deuterium exchange mass spectrometry (HDX-MS) for 0.5, 5, 50, 500 and 50,000 seconds, and compared them to each other (Fig. 2, Extended Data Figs. 1d–f, 4, 5, Supplementary Data 1). We achieved excellent peptide coverage (89, 87 and 80% for SMCR8, C9orf72 and WDR41, respectively), and consistent patterns were observed across experimental time points. Several regions in SMCR8—including the N-terminal 54 residues, and residues V104–V118, E212–I230, P257–F315, V378–I714 and

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, CA, USA. [3]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ✉e-mail: jimhurley@berkeley.edu

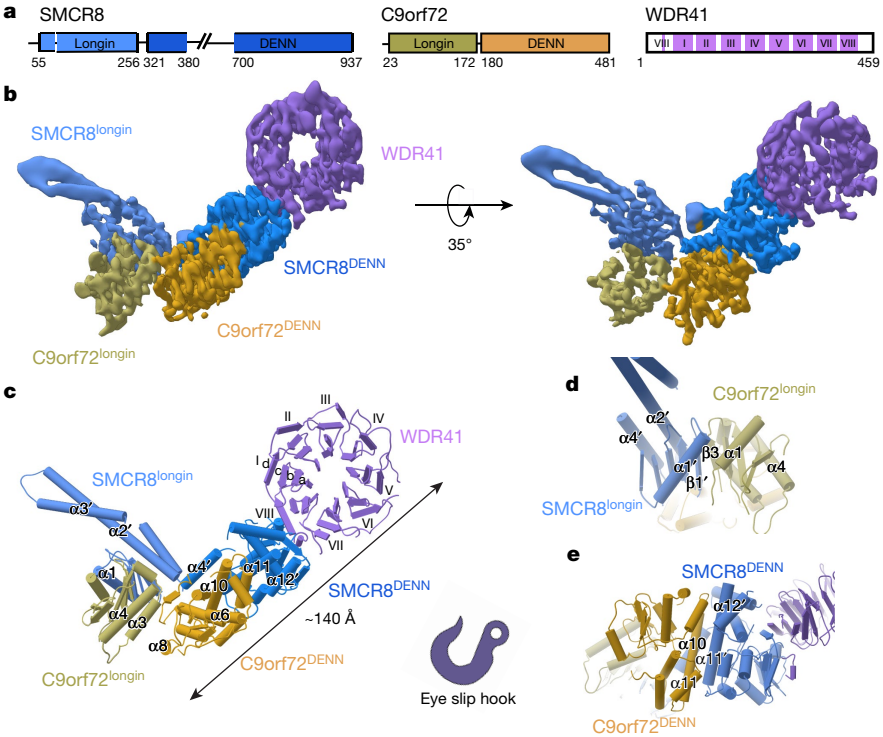


Fig. 1 | Cryo-EM structure of the C9orf72–SMCR8–WDR41 complex. **a**, Schematic of the domain structure of the C9orf72–SMCR8–WDR41 complex. **b**, **c**, Cryo-EM density map (local filter map, *b* factor, −50 Å^2) (**b**) and the refined coordinates (**c**) of the complex shown as pipes and planks for α-helices and β-sheets, respectively. The domains are colour-coded as follows: longin domain of SMCR8 (SMCR8^{longin}), cornflower blue; DENN domain of SMCR8 (SMCR8^{DENN}), Dodger blue; longin domain of C9orf72 (C9orf72^{longin}), olive; DENN domain of C9orf72 (C9orf72^{DENN}), goldenrod; and WDR41, medium purple. **d**, **e**, Organizations of SMCR8^{longin}–C9orf72^{longin} (**d**) and SMCR8^{DENN}–C9orf72^{DENN} (**e**) arrangement.

V788–Y806—showed more than 50% deuterium uptake at 0.5 seconds, which indicates these regions are intrinsically disordered regions—consistent with sequence-based predictions. Nearly all of C9orf72 was protected from exchange, except for the N-terminal 21 residues and the C terminus. For WDR41, the N-terminal 24 residues, and the loops that connect blades II and III (R128–C131), blades V and VI (R260–D270 and L277–I284), the internal loop of blade VII and the loop connecting to blade VIII (R352–L357 and M369–E396) were flexible.

Difference heat maps for C9orf72 and SMCR8 (Fig. 2a, b) showed that, in the presence of WDR41, regions of the DENN domain of SMCR8—including K363–F372 (M1 region), P763–Q778 (M2 region), S729–V735 (M3 region), T807–D811 (M4 region) and the C-terminal K910–Y935 (M5 region)—were protected from exchange (Figs. 2, 3, Extended Data Figs. 4–6), consistent with the structure. There was no substantial exchange in C9orf72, with the exception of K388–R394 (the M1 region of C9orf72) (Figs. 2, 3). We mutagenized the regions that showed protection against exchange, and tested these mutants in co-expression and pulldown experiments (Fig. 2c, d; see 'Protein expression and purification' in Methods for details of the mutants). Except for the helical linker mutant in the M1 region of SMCR8, the SMCR8 mutants abolished the interaction with WDR41. When WDR41 did not pull down SMCR8 mutants, wild-type C9orf72 was not detected either. This confirms the structural finding that SMCR8 bridges the other two components. Because alterations of the M1 region of C9orf72 did not prevent interaction with SMCR8–WDR41, we concluded that this region was protected by a conformational change induced upon WDR41 binding, consistent with the lack of direct interaction in the cryo-EM structure. The interface between SMCR8 and C9orf72 is extensive, mediated by longin–longin and DENN–DENN dimerization (Fig. 1d, e). Substitutions in the C9orf72(F397E/T411W) double mutant disrupt the interaction with SMCR8, as shown by co-expression and pulldown experiments (Extended Data Fig. 7a, b). The cryo-EM structure showed that SMCR8 bound to blade VIII and the C-terminal helix of WDR41 (Fig. 3a, Extended Data Fig. 6). The pulldown experiment showed that the N-terminal residues E35–K40 of blade VIII and the C-terminal helix S442–V459 are required for SMCR8 binding (Extended Data Fig. 7c). Collectively, the HDX-MS and mutational results corroborate the structural interpretation.

WDR41 is responsible for the reversible targeting of C9orf72–SMCR8 to lysosomes under conditions of nutrient depletion[14]. WDR41, in turn, binds to lysosomes via PQLC2[15]. We cotransfected DNA encoding green fluorescent protein-tagged SMCR8 (GFP–SMCR8), C9orf72, WDR41 and PQLC2 tagged with monomeric red fluorescent protein (PQLC2–mRFP) in HEK293A cells. SMCR8 clustered on PQLC2-positive lysosomes in conditions of amino acid depletion and was diffusely localized in the cytosol upon refeeding (Fig. 3b), consistent with previous reports[14,15]. SMCR8 mutants deficient in WDR41 binding in vitro did not colocalize with PQLC2-positive lysosomes, but rather were diffusely localized in the cytosol even under amino acid-starved conditions (Fig. 3b, c). These findings confirm that the WDR41-binding site on SMCR8 as mapped by cryo-EM and HDX-MS is responsible for the lysosomal localization of the complex under conditions of amino acid starvation.

The structure showed that the longin domain of SMCR8 forms a heterodimer with the longin domain of C9orf72 in the same manner as NPRL2–NPRL3 of the GATOR1 complex[19] and FLCN–FNIP2 in the lysosomal folliculin complex[20,21]. The NPRL2 and FLCN subunits of these complexes are GAPs for the lysosomal small GTPases RAGA[22] and RAGC[23], respectively. Structure-based alignment of SMCR8 with FLCN and NPRL2 showed that they shared a conserved arginine finger residue[20,21,24] (Fig. 4a),which corresponds to SMCR8 R147. This arginine residue is exposed on the protein surface near the centre of a large concave surface, which appears suitable for binding a small GTPase (Extended Data Fig. 8). Using a tryptophan fluorescence and high-performance liquid chromatography (HPLC)-based assay, we assayed C9orf72–SMCR8

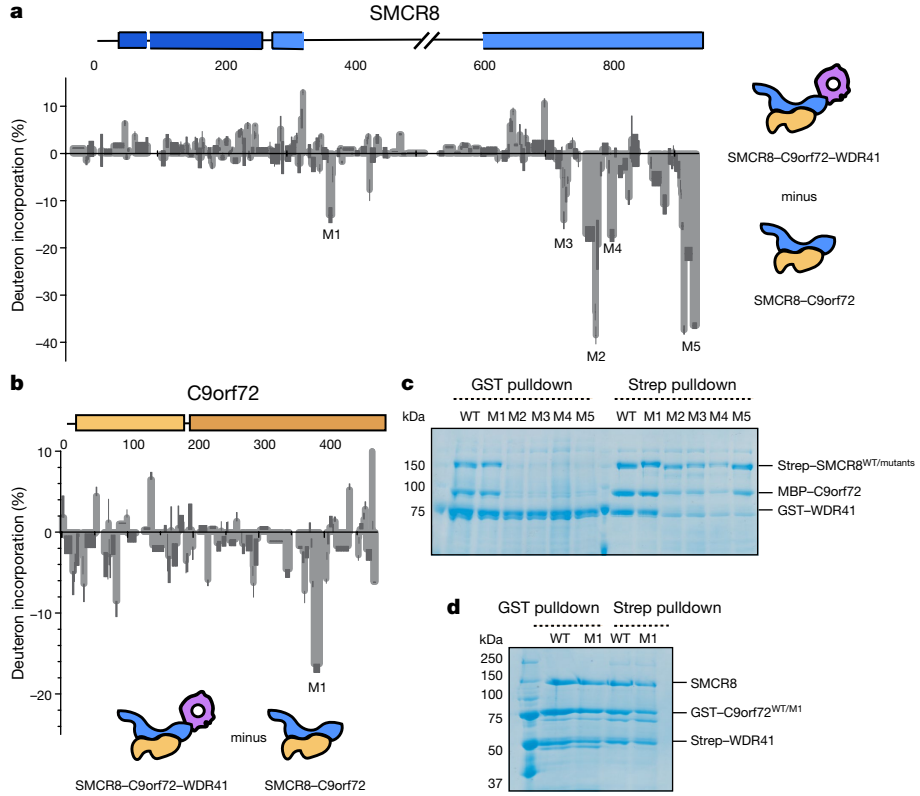


Fig. 2 | HDX-MS of C9orf72–SMCR8 in the absence of WDR41. a, Difference plot of percentage of deuteron incorporation of SMCR8 in the heterotrimer versus the dimer, at the 5-s time point. **b**, Difference plot of percentage of deuteron incorporation of C9orf72 in the heterotrimer versus the dimer, at the 0.5-s time point. **c**, Co-expression and pulldown experiment of Strep-tagged SMCR8 mutants with wild-type MBP–C9orf72 and GST–WDR41. M1–5, mutants of the corresponding M1–5 regions of SMCR8; Strep–SMCR8$^{WT/mutants}$: WT, wild type; mutants, mutants of the M1–5 regions of SMCR8. **d**, Co-expression and pulldown experiments of GST–C9orf72 mutant with wild-type untagged SMCR8 and Strep–WDR41. M1, mutant of the M1 region of C9orf72; GST–C9orf72$^{WT/M1}$: WT, wild type; M1, mutant of the M1 region of C9orf72. The pulldown experiments were repeated at least twice with similar results (**c**, **d**).

for GAP activity with respect to RAGA or RAGC and found none detectable (Extended Data Fig. 9a, b, d). We also assayed for GAP activity with respect to RAB1A[17] and the late endosomal RAB7A[16], and—again—activity was undetectable (Extended Data Fig. 9a, b, d).

It has previously been reported that C9orf72 interacts with the small GTPases ARF1 and ARF6[25] in neurons[5]. We found that C9orf72–SMCR8–WDR41 was an efficient GAP for ARF1 on the basis of both tryptophan fluorescence and HPLC-based assays (Fig. 4). The ARF1(Q71L) GTP-locked mutant had no activity (Fig. 4b, Extended Data Fig. 10); nor did the version of the complex that contained the SMCR8(R147A) finger mutant. FLCN–FNIP2 and GATOR1 had no GAP activity towards ARF1. C9orf72–SMCR8 was as active as C9orf72–SMCR8–WDR41, consistent with the location of WDR41 on the opposite side of the complex from R147. C9orf72–SMCR8–WDR41 has activity against the other members of the ARF family, ARF5 and ARF6 (Extended Data Fig. 9a, c, d)—but not against the lysosomal ARF-like proteins ARL8A and ARL8B (Extended Data Fig. 9a, b, d). These observations clarify the nature of the reported C9orf72–ARF interaction by showing that the role of C9orf72 is to stabilize a complex with SMCR8, which is—in turn—an efficient and selective GAP for ARF GTPases.

RAB5A[26], RAB7A[26], RAB8A[8] and RAB39B[8,12] have all previously been reported to be guanine nucleotide exchange factor (GEF) substrates of C9orf72. We tested the activity of the purified complex with respect to these RAB proteins and another putative C9orf72 interactor, RAB1A[17]. Compared to a RABEX5 and RAB5A positive control, no exchange was observed on any of these upon addition of C9orf72–SMCR8–WDR41 (Extended Data Fig. 11a, b). The structure of RAB35 in complex with the GEF DENND1B[27] was previously used as a basis for modelling[26]. In comparing C9orf72 in our structure with the structure of DENND1B in complex with RAB35[27], we found that the alignment of the longin domains showed that RAB35 collides with the longin domain of SMCR8, and superimposition of DENN domains indicated that RAB35 collides with the longin domain of C9orf72—consistent with our result that C9orf72–SMCR8 does not have DENND1B-like GEF activity (Extended Data Fig. 11c).

These data shed light on the normal function of C9orf72, which is thought to contribute to neuronal loss-of-function in ALS and FTD[6]. The structure shows that C9orf72 is the central component of its complex with SMCR8. The longin and DENN domains of SMCR8 flank, and are stabilized by, C9orf72. SMCR8 contains the binding site for WDR41, which is responsible for lysosomal localization during amino acid starvation. C9orf72–SMCR8 belongs to the same class of double-longin-domain GAP complexes as GATOR1[19] and FLCN–FNIP2[20,21]. Unlike GATOR1 and FLCN–FNIP2, C9orf72–SMCR8 is inactive against RAG GTPases but is active against ARF GTPases. The GAP active site is located at the opposite end of the complex from the lysosomal targeting site on WDR41.

Our in vitro observation that C9orf72–SMCR8 and C9orf72–SMCR8–WDR41 have comparable GAP activities suggests that—in cells—C9orf72–SMCR8 may regulate ARF GTPases both in full nutrient conditions, when the complex is primarily localized in the cytosol, and under conditions of amino acid starvation, when it relocalizes to the lysosomal membrane via interactions between WDR41 and PQLC2. However, additional factors could limit or augment the ARF GAP activity of C9orf72–SMCR8 in either condition, and restrict or enhance access to the GTP-bound ARF substrate. ARF proteins are not observed on lysosomes, and their closest lysosomal cousins (ARL8A and ARL8B) are not substrates for C9orf72–SMCR8. Thus, sequestration of C9orf72–SMCR8–WDR41 on lysosomes could prevent it from

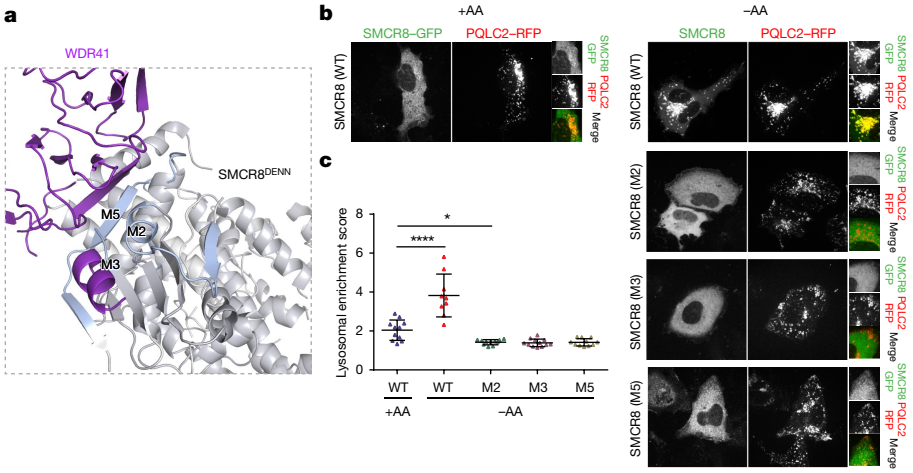


Fig. 3 | SMCR8 mutants do not localize on lysosomes. a, The HDX uptake difference at 0.5 s was mapped on C9orf72–SMCR8. Close-up view of the SMCR8–WDR41 interface, highlighting the SMCR8 mutants. **b**, SMCR8–PQLC2 lysosome colocalization experiment in cells that express the indicated SMCR8 constructs under the indicated nutrient conditions. –AA, cells starved for amino acids for 1 h; +AA, cells starved for amino acids for 1 h, and then restimulated with amino acids for 10 min. The experiment was repeated at least three times independently with similar results. **c**, Quantification of SMCR8 lysosomal enrichment score for fluorescence images shown in **b**. Data are mean ± s.d. From left to right, $n = 11, 9, 11, 12$ and 11 cells were quantified for each condition. $^{*}P < 0.05$, $^{****}P < 0.0001$, evaluated by one-way analysis of variance.

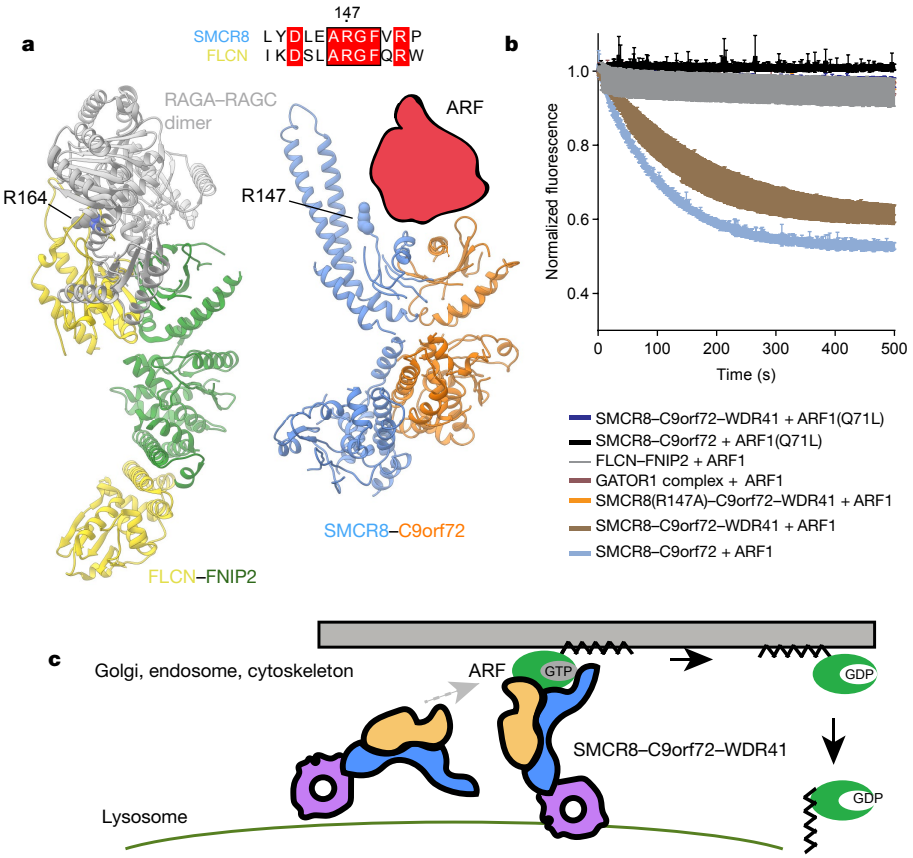


Fig. 4 | C9orf72–SMCR8 is a GAP for ARF proteins. a, Structure comparison of FNIP2–FLCN and C9orf72–SMCR8, which suggests a potential binding site for substrates. The conserved arginine residue is shown in spherical representation. **b**, Tryptophan fluorescence GTPase signal was measured for wild-type ARF1 or ARF1(Q71L) before and after the addition of the C9orf72–SMCR8–WDR41, C9orf72–SMCR8(R147A)–WDR41, C9orf72–SMCR8, FLCN–FNIP2 or GATOR1 complex. The fluorescence signal upon addition of the GAP was normalized to 1 for each experiment. Plots are mean ± s.d. of triplicate technical experiments. **c**, Model for ARF protein family activation by C9orf72–SMCR8–WDR41.

regulating the ARF proteins in *cis* under unfavourable metabolic conditions. Alternatively, C9orf72–SMCR8–WDR41 could act in *trans* on ARF proteins bound to the membrane of a compartment other than the lysosome. ARF GTPases are found on the Golgi, endosomes, plasma membrane and cytoskeleton, and in the cytosol[25], and function on membranes in their active GTP-bound form. C9orf72 can associate

with endosomes[6,16,28] and the cytoskeleton[5], which could be loci of the ARF substrate of C9orf72–SMCR8. The potential *trans* GAP activity of C9orf72–SMCR8–WDR41 versus endosomal or cytoskeletal ARF would be facilitated by its elongated structure and the distal positioning of the GAP and lysosomal localization sites (Fig. 4c).

Haploinsufficient GAP activity for ARF GTPases could contribute to ALS and FTD in several ways. Defects in actin dynamics in neurons could contribute to problems with endosomal transport[5]. Indeed, several studies connect C9orf72 to endosomal sorting[6,16,28], a process in which the role of ARF proteins is well-established[25]. It has previously been reported that ARF1 promotes mTORC1 activation[29], so the GAP function of C9orf72–SMCR8 with respect to ARF GTPases could explain how this complex antagonizes mTORC1[18]. mTORC1 negatively regulates autophagy, and thus the ARF1–mTORC1 connection could explain how haploinsufficient C9orf72 leads to a decrease in autophagy—which has, in turn, previously been linked to multiple neurodegenerative diseases[30]. While this Article was under review, the cryo-EM structure of a dimeric form of the C9orf72–SMCR8–WDR41 complex was reported and proposed to serve as a GAP for RAB8A and RAB11A[31]. The relative roles of GAP activity with respect to different small GTPases in normal function and disease remain to be determined. The structural and in vitro biochemical data reported here, and previously[31], provide a framework and a foothold for understanding how the normal functions of C9orf72 relate to lysosomal signalling, autophagy and neuronal survival.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2633-x.

1. Majounie, E. et al. Frequency of the *C9orf72* hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurol*. **11**, 323–330 (2012).
2. DeJesus-Hernandez, M. et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of *C9ORF72* causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
3. Renton, A. E. et al. A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
4. O'Rourke, J. G. et al. *C9orf72* is required for proper macrophage and microglial function in mice. *Science* **351**, 1324–1329 (2016).
5. Sivadasan, R. et al. C9ORF72 interaction with cofilin modulates actin dynamics in motor neurons. *Nat. Neurosci*. **19**, 1610–1618 (2016).
6. Shi, Y. et al. Haploinsufficiency leads to neurodegeneration in *C9ORF72* ALS/FTD human induced motor neurons. *Nat. Med*. **24**, 313–325 (2018).
7. Zhang, D., Iyer, L. M., He, F. & Aravind, L. Discovery of novel DENN proteins: implications for the evolution of eukaryotic intracellular membrane structures and human disease. *Front. Genet*. **3**, 283 (2012).
8. Sellier, C. et al. Loss of C9ORF72 impairs autophagy and synergizes with polyQ Ataxin-2 to induce motor neuron dysfunction and cell death. *EMBO J*. **35**, 1276–1297 (2016).
9. Sullivan, P. M. et al. The ALS/FTLD associated protein C9orf72 associates with SMCR8 and WDR41 to regulate the autophagy-lysosome pathway. *Acta Neuropathol. Commun*. **4**, 51 (2016).
10. Amick, J., Roczniak-Ferguson, A. & Ferguson, S. M. C9orf72 binds SMCR8, localizes to lysosomes, and regulates mTORC1 signaling. *Mol. Biol. Cell* **27**, 3040–3051 (2016).
11. Ugolino, J. et al. Loss of C9orf72 enhances autophagic activity via deregulated mTOR and TFEB signaling. *PLoS Genet*. **12**, e1006443 (2016).
12. Yang, M. et al. A C9ORF72/SMCR8-containing complex regulates ULK1 and plays a dual role in autophagy. *Sci. Adv*. **2**, e1601167 (2016).
13. Jung, J. et al. Multiplex image-based autophagy RNAi screening identifies SMCR8 as ULK1 kinase activity and gene expression regulator. *eLife* **6**, e23063 (2017).
14. Amick, J., Tharkeshwar, A. K., Amaya, C. & Ferguson, S. M. WDR41 supports lysosomal response to changes in amino acid availability. *Mol. Biol. Cell* **29**, 2213–2227 (2018).
15. Amick, J., Tharkeshwar, A. K., Talaia, G. & Ferguson, S. M. PQLC2 recruits the C9orf72 complex to lysosomes in response to cationic amino acid starvation. *J. Cell Biol*. **219**, e20190676 (2020).
16. Farg, M. A. et al. C9ORF72, implicated in amytrophic lateral sclerosis and frontotemporal dementia, regulates endosomal trafficking. *Hum. Mol. Genet*. **23**, 3579–3595 (2014).
17. Webster, C. P. et al. The C9orf72 protein interacts with Rab1a and the ULK1 complex to regulate initiation of autophagy. *EMBO J*. **35**, 1656–1676 (2016).
18. Lan, Y., Sullivan, P. M. & Hu, F. SMCR8 negatively regulates AKT and MTORC1 signaling to modulate lysosome biogenesis and tissue homeostasis. *Autophagy* **15**, 871–885 (2019).
19. Shen, K. et al. Architecture of the human GATOR1 and GATOR1–Rag GTPases complexes. *Nature* **556**, 64–69 (2018).
20. Lawrence, R. E. et al. Structural mechanism of a Rag GTPase activation checkpoint by the lysosomal folliculin complex. *Science* **366**, 971–977 (2019).
21. Shen, K. et al. Cryo-EM structure of the human FLCN–FNIP2–Rag–Ragulator complex. *Cell* **179**, 1319–1329 (2019).
22. Bar-Peled, L. et al. A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science* **340**, 1100–1106 (2013).
23. Tsun, Z. Y. et al. The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Mol. Cell* **52**, 495–505 (2013).
24. Shen, K., Valenstein, M. L., Gu, X. & Sabatini, D. M. Arg78 of Nprl2 catalyzes GATOR1-stimulated GTP hydrolysis by the Rag GTPases. *J. Biol. Chem*. **294**, 2970–2975 (2019).
25. Sztul, E. et al. ARF GTPases and their GEFs and GAPs: concepts and challenges. *Mol. Biol. Cell* **30**, 1249–1271 (2019).
26. Iyer, S., Subramanian, V. & Acharya, K. R. C9orf72, a protein associated with amyotrophic lateral sclerosis (ALS) is a guanine nucleotide exchange factor. *PeerJ* **6**, e5815 (2018).
27. Wu, X. et al. Insights regarding guanine nucleotide exchange from the structure of a DENN-domain protein complexed with its Rab GTPase substrate. *Proc. Natl Acad. Sci. USA* **108**, 18672–18677 (2011).
28. Corrionero, A. & Horvitz, H. R. A C9orf72 ALS/FTD ortholog acts in endolysosomal degradation and lysosomal homeostasis. *Curr. Biol*. **28**, 1522–1535 (2018).
29. Jewell, J. L. et al. Differential regulation of mTORC1 by leucine and glutamine. *Science* **347**, 194–198 (2015).
30. Nixon, R. A. The role of autophagy in neurodegenerative disease. *Nat. Med*. **19**, 983–997 (2013).
31. Tang, D. et al. Cryo-EM structure of C9ORF72–SMCR8–WDR41 reveals the role as a GAP for Rab8a and Rab11a. *Proc. Natl Acad. Sci. USA* **117**, 9876–9883 (2020).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification

Synthetic genes encoding SMCR8 were amplified by PCR and cloned into the pCAG vector coding for an N-terminal twin Strep–Flag tag using KpnI and XhoI restriction sites. The pCAG vector encoding an N-terminal GST followed by a TEV restriction site or uncleaved MBP tag was used for expression of C9orf72. WDR41 was cloned into pCAG vector without a tag or with a GST tag for pulldown experiments. For the mutations of SMCR8 identified from HDX experiments, the M1 region of SMCR8 (K363–F372) was mutated to MSDYDIPTTE, which is a 10-residue linker derived from the pETM11 vector. For lysosome localization experiments, the M2 region of SMCR8 (P771–Q778 or K762–L782) was mutated to GGKGSGGS. Mutants of the M3 (S729–V735) and M4 (T807–D811) regions of SMCR8 were made by mutating these regions to GGKGSGG and GGKGS, respectively. The mutant of the M5 region of SMCR8 was made by truncation after residue K910. The M1 region of C9orf72 (K388–L393) was mutated to polyalanine. The SMCR8 arginine finger mutation (R147A), C9orf72(F397E) and C9orf72(T411W) mutants were made using two-step PCR and cloned into the expression vector.

HEK293-GnTi cells adapted for suspension were grown in Freestyle medium supplemented with 1% FBS and 1% antibiotic–antimycotic at 37 °C, 80% humidity, 5% CO_2 and shaking at 140 rpm. Once the cultures reached 1.5–2 million cells ml^{-1} in the desired volume, they were transfected as followed. For a 1-l transfection, 3 ml PEI (1 mg ml^{-1}, pH 7.4, Polysciences) was added to 50 ml hybridoma medium (Invitrogen) and 1 mg of total DNA (isolated from transformed *Escherichia coli* XL10-gold) in another 50 ml hybridoma medium. One mg of transfection DNA contained an equal mass ratio of C9orf72-complex expression plasmids. PEI was added to the DNA, mixed and incubated for 15 min at room temperature. One hundred ml of the transfection mix was then added to each 1-l culture. Cells were collected after 3 days.

Cells were lysed by gentle rocking in lysis buffer containing 50 mM HEPES, pH 7.4, 200 mM NaCl, 2 mM $MgCl_2$, 1% (v/v) Triton X-100, 0.5 mM TCEP, protease inhibitors (AEBSF, leupeptin and benzamidine) and supplemented with phosphatase inhibitors (50 mM NaF and 10 mM β-glycerophosphate) at 4 °C. Lysates were clarified by centrifugation (15,000*g* for 40 min at 4 °C) and incubated with 5 ml glutathione sepharose 4B (GE Healthcare) for 1.5 h at 4 °C with gentle shaking. The glutathione sepharose 4B matrix was applied to a gravity column, washed with 100 ml wash buffer (20 mM HEPES, pH 7.4, 200 mM NaCl, 2 mM $MgCl_2$ and 0.5 mM TCEP), and purified complexes were eluted with 40 ml wash buffer containing 50 mM reduced glutathione. Eluted complexes were treated with TEV protease at 4 °C overnight. TEV-treated complexes were purified to homogeneity by injection on Superose 6 10/300 (GE Healthcare) column that was pre-equilibrated in gel filtration buffer (20 mM HEPES, pH 7.4, 200 mM NaCl, 2 mM $MgCl_2$ and 0.5 mM TCEP). For long-term storage, fractions from the gel filtration chromatography were frozen using liquid nitrogen and kept at −80 °C. C9orf72–SMCR8 and C9orf72–SMCR8–WDR41 were expressed and purified using the same protocol (Supplementary Fig. 1).

For expression of human His_6-tagged ARF1 (residues E17–K181), ARF1(Q71L), ARF5(Q17–Q180), ARF6(R15–S175), ARF6(Q67L), His_6–RAB1A, His_6–ARL8A(E20–S186), His_6–ARL8B(E20–S186), His_6–RAB39B and bovine His_6–RABEX5 helix bundle–Vps9 domain (S133–E398), plasmids were transformed into *E. coli* BL21 DE3 star cells and induced with 0.5 mM IPTG at 18 °C overnight. The cells were lysed in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 2 mM $MgCl_2$, 5 mM imidazole, 0.5 mM TCEP and 1 mM PMSF by ultrasonication. The lysate was centrifuged at 15,000*g* for 30 min. The supernatant was loaded into Ni-NTA resin and washed with 20 mM imidazole and eluted with 300 mM imidazole. The eluate was further purified on a Superdex 75 10/300 (GE Healthcare) column equilibrated in 20 mM HEPES, pH 7.4, 200 mM NaCl, 2 mM $MgCl_2$ and 0.5 mM TCEP. RAG, FLCN–FNIP2 and GATOR1 complex were purified as previously described[20]. GST-tagged human RAB7A, or RAB5A (*Canis familiaris*), was expressed in the same conditions as above and purified with GST resin, eluted in 50 mM reduced glutathione buffer and applied on Superdex 200 column. Twin Strep–Flag tag RAB8A was expressed in HEK293-GnTi cells and purified by Strep resin and eluted in 10 mM desthiobiotin buffer. The eluted protein was applied on Superdex 75 10/300 column.

Hydrogen–deuterium exchange experiment

Sample quality was assessed by SDS–PAGE before each experiment. Amide hydrogen exchange mass spectrometry was initiated by a 20-fold dilution of 10 μM C9orf72–SMCR8–WDR41 or C9orf72–SMCR8 into 95 μl D_2O buffer containing 20 mM HEPES pH (pD 8.0), 200 mM NaCl, 1 mM TCEP at 30 °C. Incubations in deuterated buffer were performed at intervals from 0.5, 5, 50, 500 and 50,000 s (0.5 s was carried out by incubating proteins with ice-cold D_2O for 5 s). All exchange reactions were carried out in triplicate or quadruplicate. Backbone amide exchange was quenched at 0 °C by the addition of ice-cold quench buffer (400 mM KH_2PO_4/H_3PO_4, pH 2.2). The 50,000-s sample served as the maximally labelled control. Quenched samples were injected onto a chilled HPLC setup with in-line peptic digestion and then eluted onto a BioBasic 5 μM KAPPA Capillary HPLC column (Thermo Fisher Scientific), equilibrated in buffer A (0.05% TFA), using 10–90% gradient of buffer B (0.05% TFA, 90% acetonitrile) over 30 min. Desalted peptides were eluted and directly analysed by an Orbitrap Discovery mass spectrometer (Thermo Fisher Scientific). The spray voltage was 3.4 kV and the capillary voltage was 37 V. The HPLC system was extensively cleaned between samples. Initial peptide identification was performed via tandem mass spectrometry experiments. A Proteome Discoverer 2.1 (Thermo Fisher Scientific) search was used for peptide identification and coverage analysis against entire complex components, with precursor mass tolerance ± 10 ppm and fragment mass tolerance of ± 0.6 Da. Mass analysis of the peptide centroids was performed using HDExaminer (Sierra Analytics), followed by manual verification of each peptide. The difference plots were prepared using Origin 6.0.

Cryo-EM grid preparation and data acquisition

The purified C9orf72–SMCR8–WDR41 complex was diluted to 0.8 μM in 20 mM HEPES pH 7.4, 2 mM $MgCl_2$ and 0.5 mM TCEP, and applied to glow-discharged C-flat (1.2/1.3, Au 300 mesh) grids. The sample was vitrified after blotting for 2 s using a Vitrobot Mark IV (FEI) with 42-s incubation, blot force 8 and 100% humidity. The complex was visualized with a Titan Krios electron microscope (FEI) operating at 300 kV with a Gatan Quantum energy filter (operated at 20-eV slit width) using a K2 summit direct electron detector (Gatan) in super-resolution counting mode, corresponding to a super-resolution pixel size of 0.5745 Å on the specimen level. In total, 3,508 movies were collected in nanoprobe mode using Volta phase plate (VPP) with defocus collected at around −60 nm. Movies consisted of 49 frames, with a total dose of 59.8 e^- per $Å^2$, a total exposure time of 9.8 s and a dose rate of 8.1 e^- per pixel per s. Data were acquired with SerialEM using custom macros for automated single-particle data acquisition. Imaging parameters for the dataset are summarized in Extended Data Table 1.

Cryo-EM data processing

Preprocessing was performed during data collection within Focus[32]. Drift, beam-induced motion and dose weighting were corrected with MotionCor2[33] using 5 × 5 patches and Fourier cropping with a factor of 0.5 after motion correction. CTF fitting and phase-shift estimation were performed using Gctf v.1.06[34], which yielded the characterized pattern of phase-shift accumulation over time for each position. The data were manually inspected and micrographs with excess ice contamination or shooting on the carbon were removed. A total of 4,810,184 particles

from 3,220 micrographs were picked using gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/) and extracted with binning 4. All subsequent classification and reconstruction steps were performed using Relion3-beta[35] or cryoSPARC v.2[36]. The particles were subjected to 3D classification ($K = 5$) using a 60 Å low-pass-filtered ab initio reference generated in cryoSPARC. Around 2.2 million particles from the two best classes were selected for 3D auto-refinement and another round of 3D classification ($K = 8$ classes, $T = 8$, E-step = 8 Å) without alignment. About 1.8 million particles from the best 6 classes were re-extracted with binning 2 and refined to 4.9 Å, and further subjected to 2D classification without alignment for removing contamination and junk particles. After another round of 3D classification ($K = 4$) with alignment, the best class was extracted and imported into cryoSPARC v.2 for another round of 2D classification. The cleaned-up 571,002 particles were subjected to CTF refinement, Bayesian polishing, and further particles at the edges were removed in Relion 3. A final set of 381,450 particles resulted in final resolution of 3.8 Å, with a measured map B-factor of −102 Å^2. More extensive 3D classification and focus classification in Relion3 did not improve the quality of the reconstruction. Local filtering and B-factor sharpening were done in cryoSPARC v.2. All reported resolutions are based on the gold-standard Fourier shell correlation (FSC) 0.143 criterion.

Atomic model building and refinement

The model of WDR41 was generated with I-Tasser[37] and used Protein Data Bank codes (PDB) 5NNZ, 2YMU, 5WLC, 4NSX and 6G6M as starting models. The model of the longin domain of C9orf72 was generated on the basis of the longin domain of NPRL2 (PDB 6CES) in Modeller[38]. The model of the DENN domain of SMCR8 was generated from Modeller and RaptorX[39] using the DENN domain of FLCN (PDB 3V42) or DENND1B (PDB 3TW8) as templates. The longin domain of SMCR8 and DENN domain of C9orf72 were generated with Phyre2[40] using longin domain of FLCN and DENN domain of FNIP2 (PDB 6NZD), respectively, as templates. Secondary structure predictions of each protein were carried out with Phyre2[40] or Psipred[41]. The models were docked into the 3D map as rigid bodies in UCSF Chimera[42]. The coordinates of the structures were manually adjusted and rebuilt in Coot[43]. The resulting models were refined using Phenix.real_space.refine in the Phenix suite with secondary structure restraints and a weight of 0.1[44,45]. Model quality was assessed using MolProbity[46] and the map-versus-model FSC (Extended Data Fig. 3c, Extended Data Table 1). Data used in the refinement excluded spatial frequencies beyond 4.2 Å to avoid over fitting. A half-map cross-validation test showed no indication of overfitting (Extended Data Fig. 3d). Figures were prepared using UCSF Chimera[42] and PyMOL v.1.7.2.1.

Live cell imaging

Eight hundred thousand HEK 293A cells were plated onto fibronectin-coated glass-bottom Mattek dishes and transfected with the indicated wild-type GFP–SMCR8 or mutants, C9orf72, WDR41 and PQLC2–mRFP constructs with transfection reagent Xtremegene. Twenty-four h later, cells were starved for amino acids for 1 h (−AA), or starved and restimulated with amino acids for 10 min (+AA). Cells in the −AA condition were transferred to imaging buffer (10 mM HEPES, pH7.4, 136 mM NaCl, 2.5 mM KCl, 2 mM $CaCl_2$, 1.2 mM $MgCl_2$) and cells in the +AA condition were transferred to imaging buffer supplemented with amino acids, 5 mM glucose and 1% dialysed FBS (+AA), and imaged by spinning-disc confocal microscopy. Lysosomal enrichment was scored as previously described[20] using a home-built MATLAB script to determine the lysosomal enrichment of GFP–SMCR8. The score was analysed for at least ten cells for each condition. The one-way analyses of variance were calculated using Prism 6 (Graphpad).

HPLC analysis of nucleotides

The nucleotides bound to small GTPases were assessed by heating the protein to 95 °C for 5 min followed by 5 min centrifugation at 16,000g. The supernatant was loaded onto a HPLC column (Eclipse XDB-C18, Agilent). Nucleotides were eluted with HPLC buffer (10 mM tetra-n-butylammonium bromide, 100 mM potassium phosphate pH 6.5, 7.5% acetonitrile). The identity of the nucleotides was compared to GDP and GTP standards.

HPLC-based GAP assay

HPLC-based GTPase assays were carried out by incubating 30 µl of GTPases (30 µM) with or without GAP complex at a 1:50 molar ratio for 30 min at 37 °C. Samples were boiled for 5 min at 95 °C and centrifuged for 5 min at 16,000g. The supernatant was injected onto an HPLC column as described in 'HPLC analysis of nucleotides'. The experiments were carried out in triplicate and one representative plot is shown.

Tryptophan-fluorescence-based GAP assay

Fluorimetry experiments were performed using a FluoroMax-4 (Horiba) instrument and a quartz cuvette compatible with magnetic stirring (Starna Cells), a path length of 10 mm, and were carried out in triplicate. The tryptophan fluorescence signal was collected using 297-nm excitation (1.5-nm slit) and 340-nm emission (20-nm slit). Experiments were performed in gel filtration buffer at room temperature with stirring. Data collection commenced with an acquisition interval of 1 s. Two µM GTPase was added to the cuvette initially. Once the signal was equilibrated, C9orf72–SMCR8–WDR41, C9orf72–SMCR8(R147A)–WDR41, C9orf72–SMCR8, FLCN–FNIP2 or GATOR1 complex was pipetted into the cuvette at a 1:10 molar ratio. Time (t) = 0 corresponds to GAP addition. The fluorescence signal upon GAP addition was normalized to 1 for each experiment. Mean ± s.d. of three replicates per conditions or one representative plot were plotted.

MantGDP loading for GEF assay

To load GTPases for the N-methylanthraniloyl (mant) fluorescence-based GEF assay, purified GTPases were diluted at least 1:10 into PBS buffer without $MgCl_2$ (10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 137 mM NaCl, 2.7 mM KCl). EDTA was added to a final concentration of 5 mM and incubated at room temperature for 10 min. A tenfold molar excess of mantGDP nucleotide (Millipore Sigma) was added to the GTPases and incubated for 30 min at room temperature. After addition of $MgCl_2$ to a final concentration of 20 mM and incubation at room temperature for 10 min, unbound nucleotides were removed by buffer exchange into gel filtration buffer using a PD-10 column (GE Healthcare).

GEF assay

GEF assays were carried out with the same instrument and cuvette as the tryptophan fluorescence assays (see 'Tryptophan-fluorescence-based GAP assay'). Mant fluorescence was collected using a 360-nm excitation (10-nm slit) and 440-nm emission (10-nm slit). Experiments were performed in gel filtration buffer at room temperature. Five hundred µl of gel filtration buffer was added to the cuvette, and after baseline equilibration, 20 µl of the respective GTPase with or without RABEX5 or C9orf72–SMCR8–WDR41 were added to a final concentration of 350 nM. After signal equilibration, the assay commenced by addition of 20 µl of GTP to a final concentration of 5 µM (about 15-fold molar excess over the respective GTPase) and fluorescence was measured in 1-s intervals for 1,400 s. All experiments were performed in triplicates. Data were baseline-subtracted and normalized to the signal immediately after GTP addition, which also is the 0-s time point in the plots. Plots are mean ± s.d. of each triplicate experiment.

Cell line authentication

Both HEK293 GnTi and HEK 293A cell lines were purchased from the UC Berkeley Cell Culture Facility, and were authenticated by short-tandem repeat analysis and confirmed to be mycoplasma-negative by nuclear staining and fluorescence microscopy screening.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The electron microscopy density map has been deposited in the Electron Microscopy Data Bank with accession number EMD-21048. Atomic coordinates for C9orf72–SMCR8–WDR41 have been deposited in the PDB with accession number 6V4U. Source data are provided with this paper.

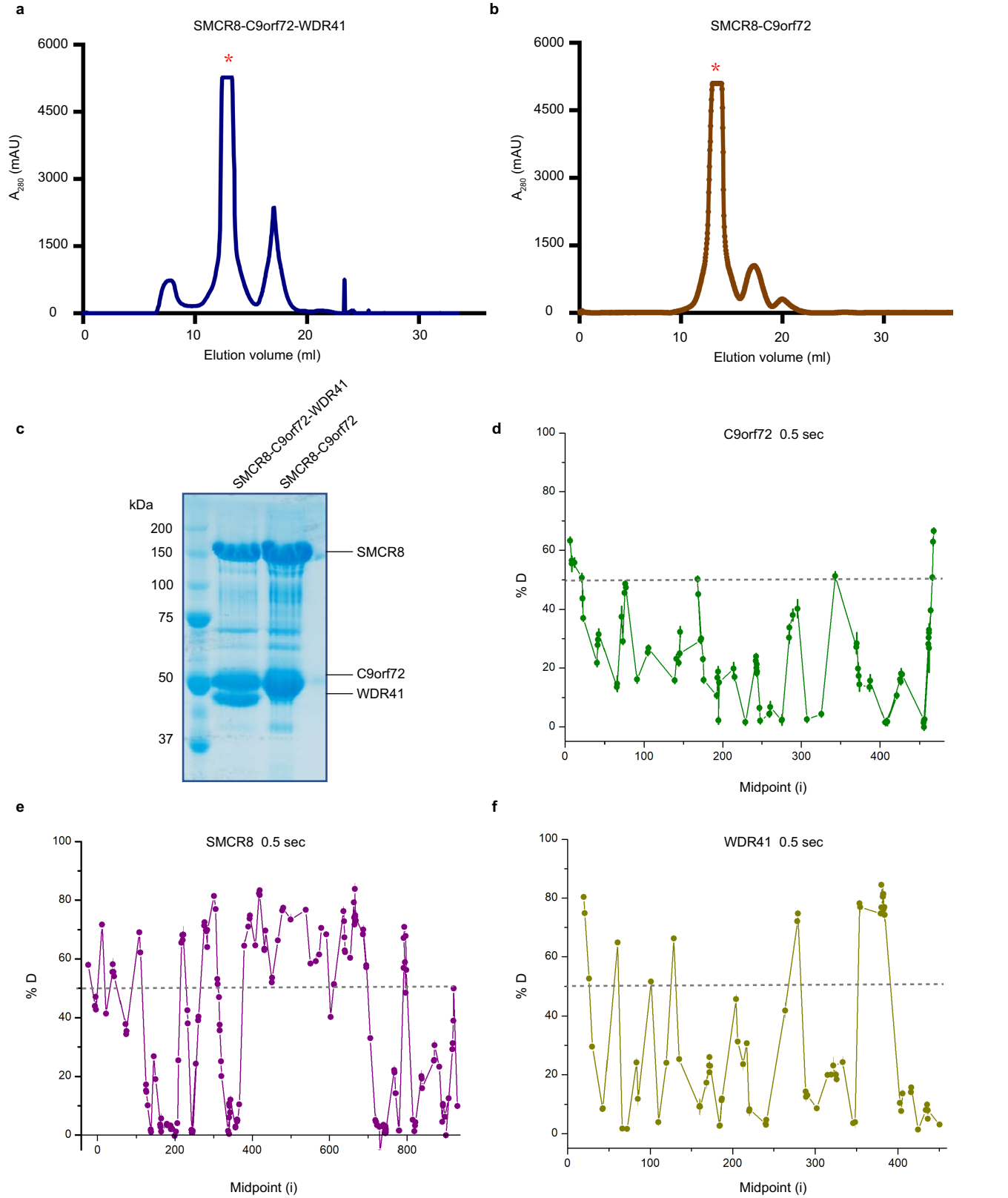
32. Biyani, N. et al. Focus: the interface between data collection and data processing in cryo-EM. *J. Struct. Biol.* **198**, 124–133 (2017).
33. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
34. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
35. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
36. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
37. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40 (2008).
38. Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **1137**, 1–15 (2014).
39. Peng, J. & Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* **79** (Suppl 10), 161–171 (2011).
40. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).
41. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000).
42. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
44. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
45. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
46. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).

Acknowledgements We thank D. Toso, K. L. Morris, V. Kasinath and P. Tobias for cryo-EM advice and support; X. Shi for HDX support; C. Behrends and G. Stjepanovic for suggestions and contributions to the early stages of the project; R. Lawrence and M. Lehmer for cell imaging advice; and X. Ren for assistance with cloning. This work was supported by NIH grants R01GM111730 (J.H.H.) and R01GM130995 (R.Z.), Department of Defense Peer Reviewed Medical Research Program Discovery Award W81XWH2010086 (J.H.H.), the Bakar Fellows program (J.H.H.), the Pew-Stewart Scholarship for Cancer Research and Damon Runyon-Rachleff Innovation Award (R.Z.), a postdoctoral fellowship from the Association for Frontotemporal Degeneration (M.-Y.S.) and an EMBO Long-Term Fellowship (S.A.F.).

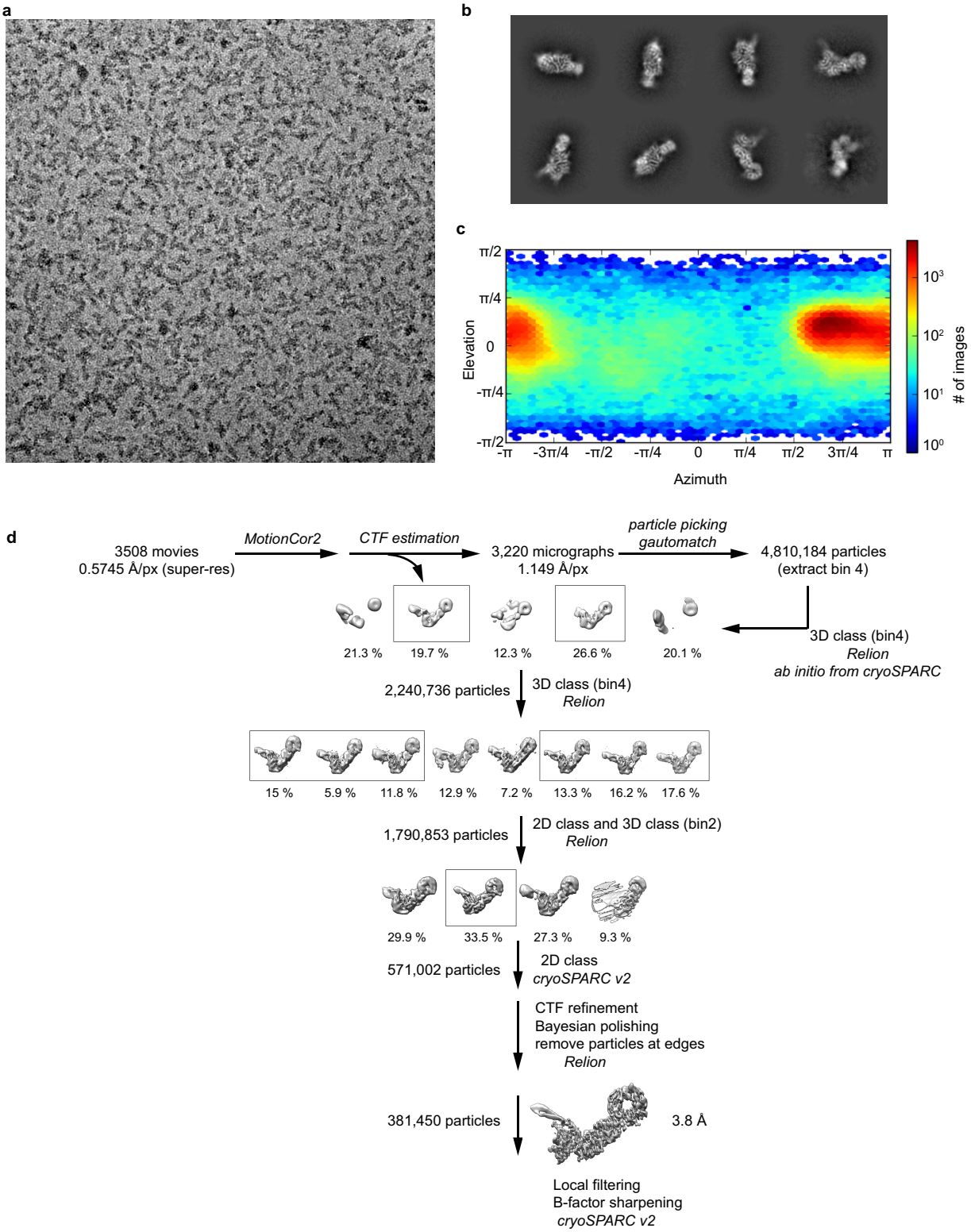
Author contributions M.-Y.S. designed and carried out all experiments, except for GEF assays and lysosome colocalization experiments, and performed all data analysis. S.A.F. performed GEF assays. R.Z. carried out lysosome colocalization experiments. M.-Y.S. and J.H.H. conceptualized the project and wrote the first draft. All authors contributed to the editing of the manuscript.

Competing interests J.H.H. is a scientific founder and receives research funding from Casma Therapeutics. R.Z. is a co-founder and stockholder in Frontier Medicines Corp.

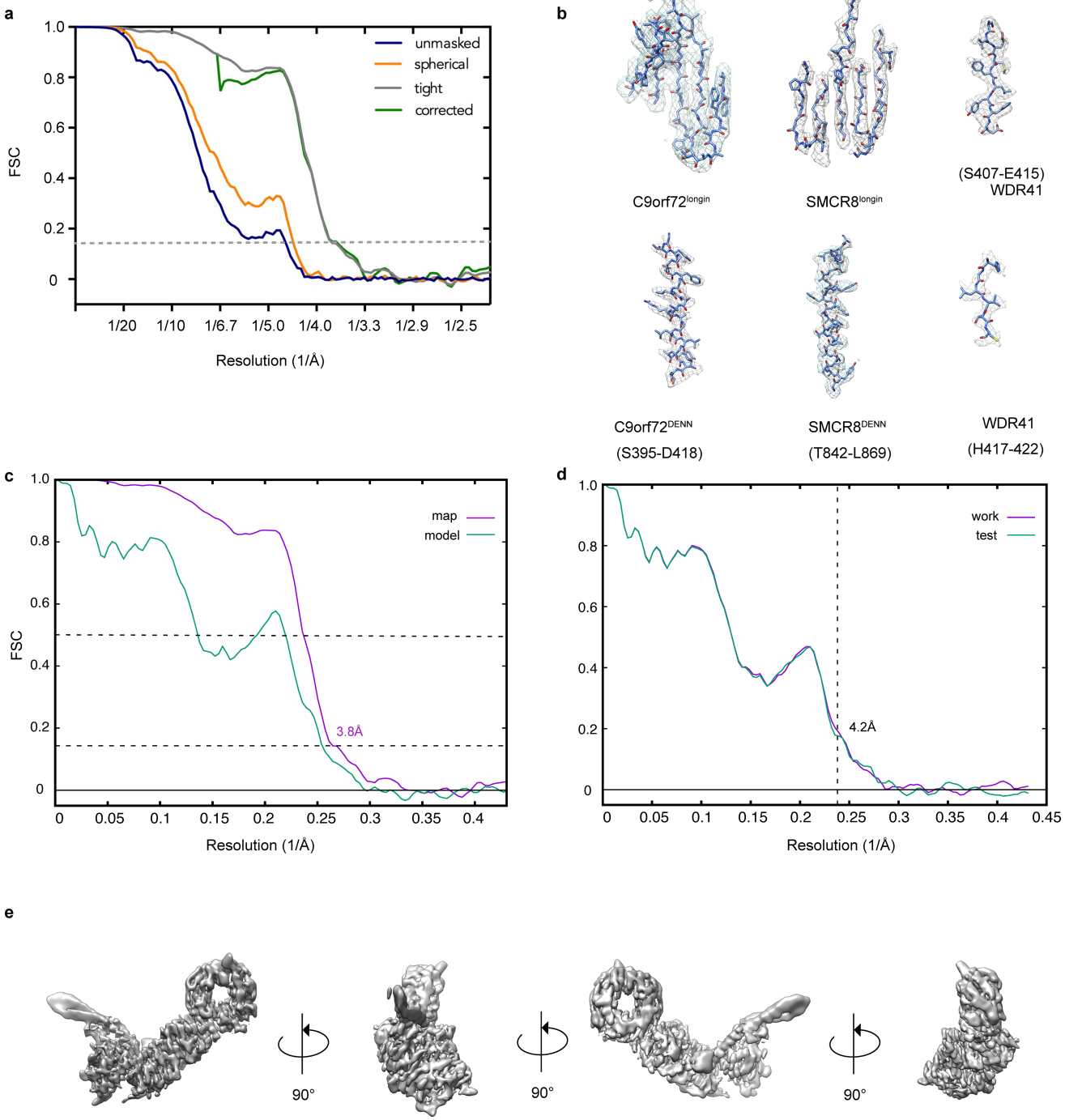
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2633-x.
Correspondence and requests for materials should be addressed to J.H.H.
Peer review information *Nature* thanks Aaron Burberry, Sjors Scheres and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



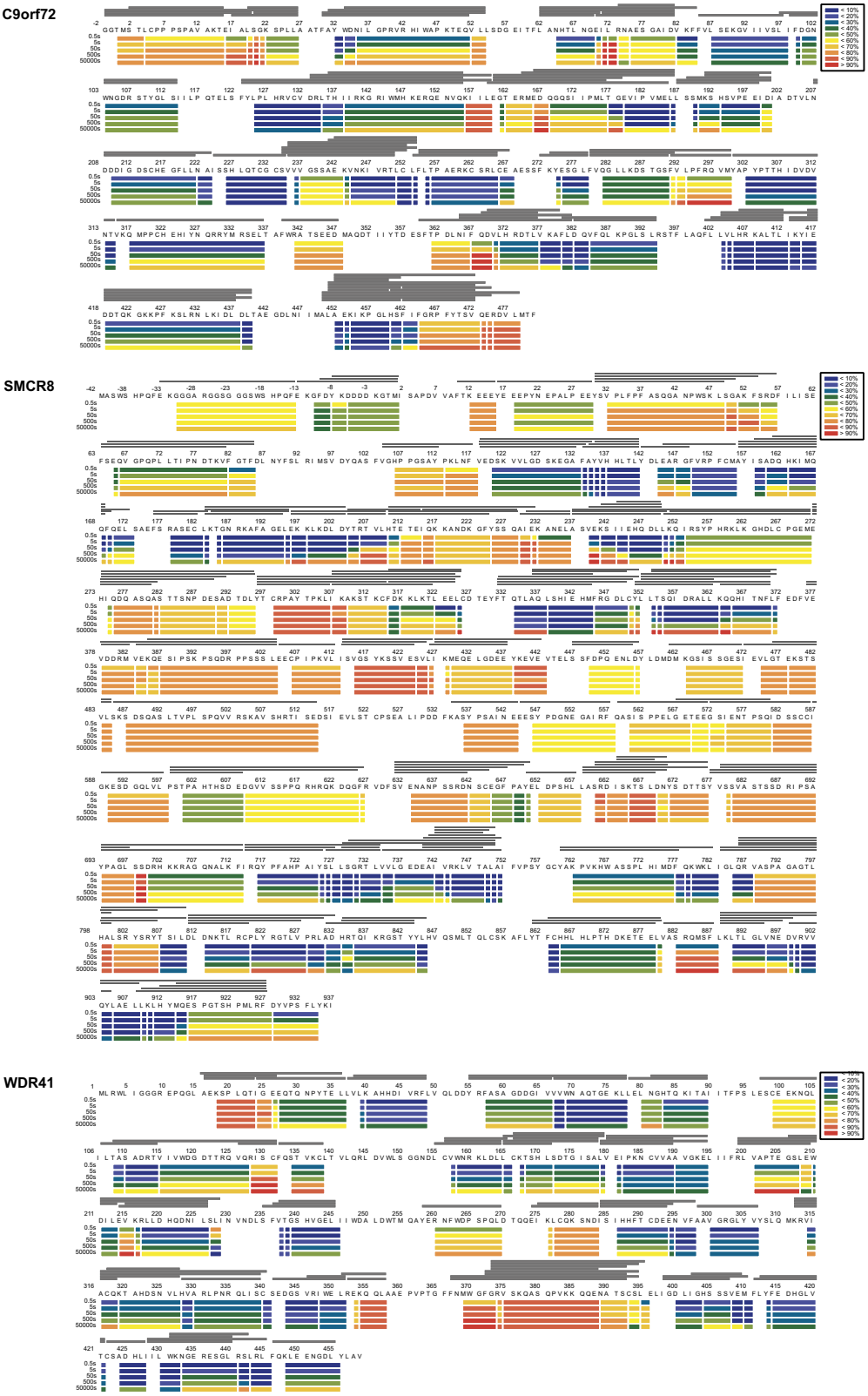
Extended Data Fig. 1 | Purification of the C9orf72–SMCR8–WDR41 and C9orf72–SMCR8 complex, as well as the HDX data for the trimer. a, The Superose-6 gel filtration elution profile for the C9orf72–SMCR8–WDR41 complex. **b**, The Superose-6 gel filtration elution profile for the C9orf72–SMCR8 complex. mAU, milli-absorbance units. **c**, The purified full-length C9orf72–SMCR8–WDR41 and C9orf72–SMCR8 complexes were analysed by SDS–PAGE. The proteins were purified at least five times with similar results (**a**–**c**). **d**–**f**, Deuterium uptake data for the C9orf72–SMCR8–WDR41 complex at the 0.5-s time point, with error bars from triplicate technical measurements. Peptides with more than 50% deuterium uptake are the flexible regions. *Y* axis represents the average per cent deuteration. *X* axis demonstrates the midpoint of a single peptic peptide.



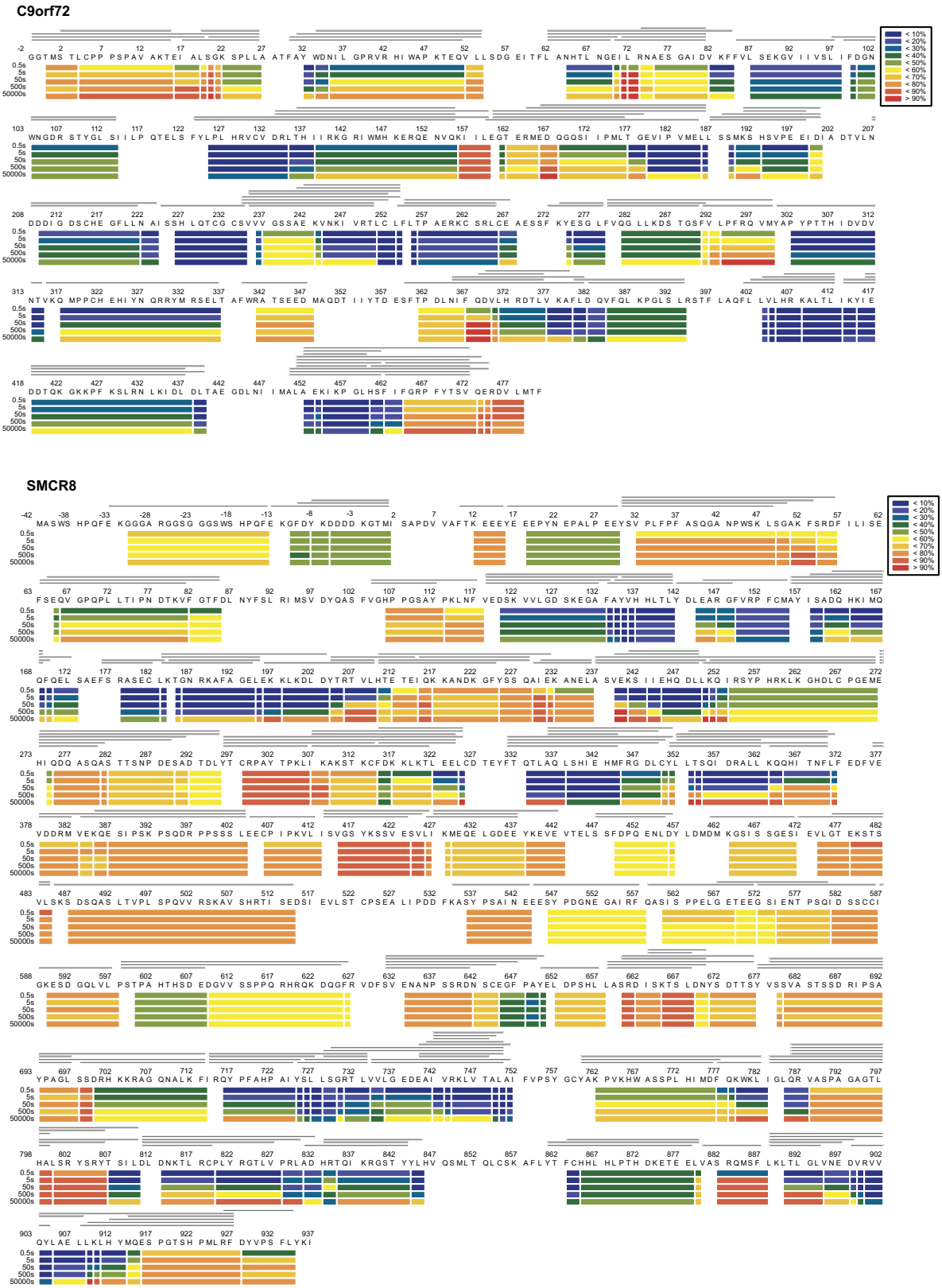
Extended Data Fig. 2 | Cryo-EM data processing. a, A representative cryo-EM micrograph of the C9orf72–SMCR8–WDR41 complex. **b**, Representative 2D classes. **c**, Orientation distribution of the aligned C9orf72–SMCR8–WDR41 particles. **d**, Image processing procedure.



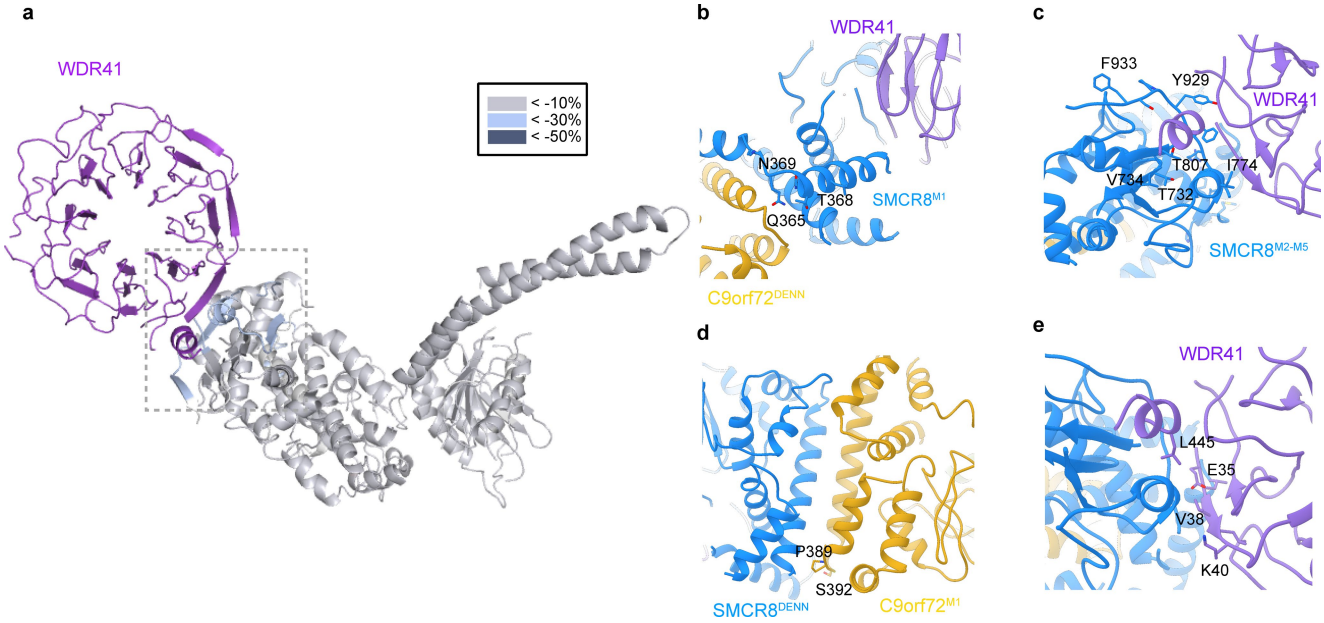
Extended Data Fig. 3 | Resolution estimation of the cryo-EM map, as well as model building and validation. a, Comparison between FSC curves. **b**, Refined coordinate model fit of the indicated region in the cryo-EM density. **c**, Refinement and map-versus-model FSC. **d**, Cross-validation of test FSC curves to assess overfitting. The refinement target resolution (4.2 Å) is indicated. **e**. Different views of the final reconstruction.



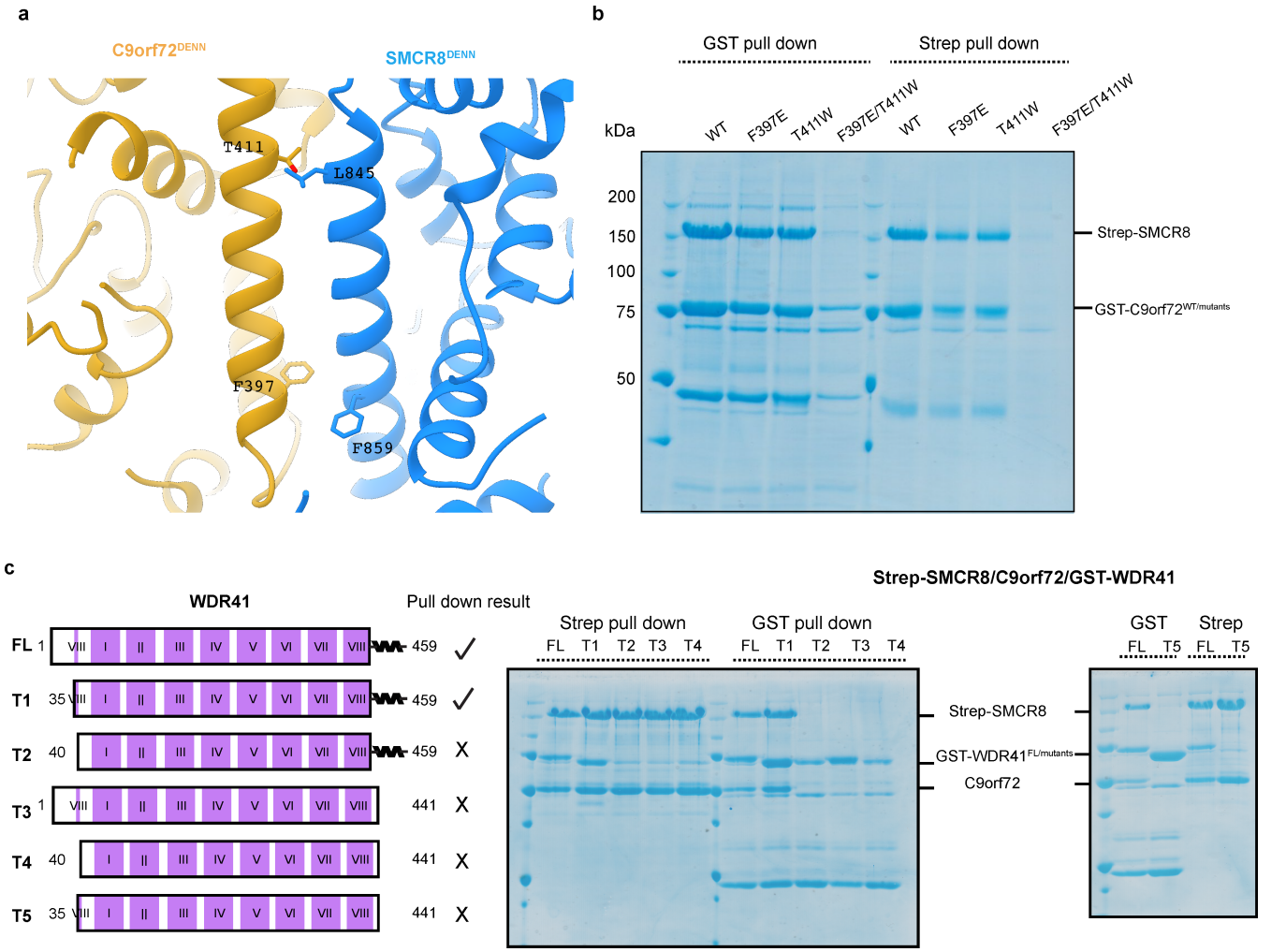
Extended Data Fig. 4 | Deuterium uptake of C9orf72–SMCR8–WDR41. HDX-MS data are shown in heat map format, in which peptides are represented by rectangular strips above the protein sequence. Absolute deuterium uptake after 0.5, 5, 50, 500 and 50,000 s are indicated by a colour gradient below the protein sequence.



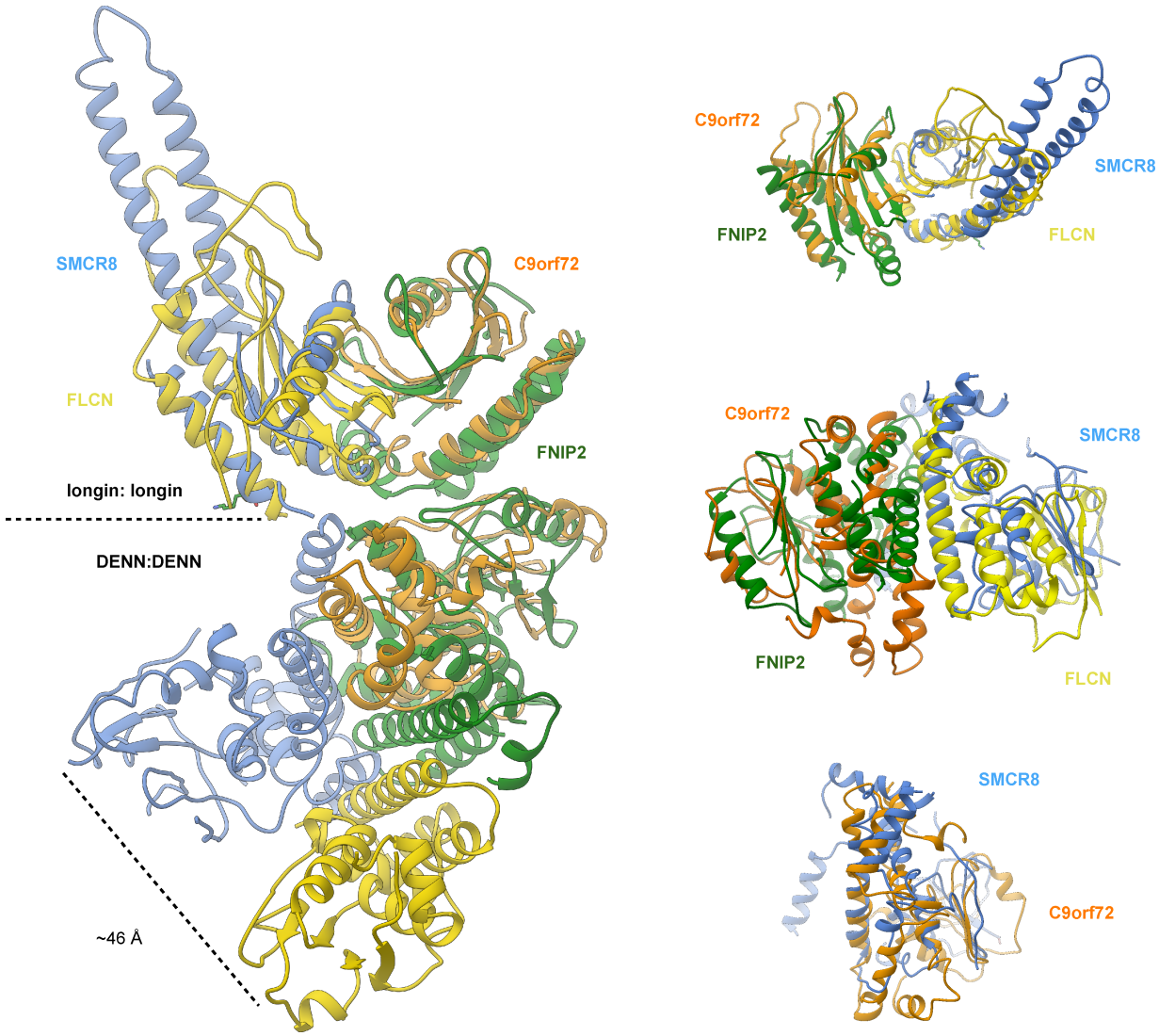
Extended Data Fig. 5 | Deuterium uptake of C9orf72–SMCR8 complex. HDX-MS data are shown in heat map format, in which peptides are represented by rectangular strips above the protein sequence. Absolute deuterium uptake after 0.5, 5, 50, 500 and 50,000 s are indicated by a colour gradient below the protein sequence.



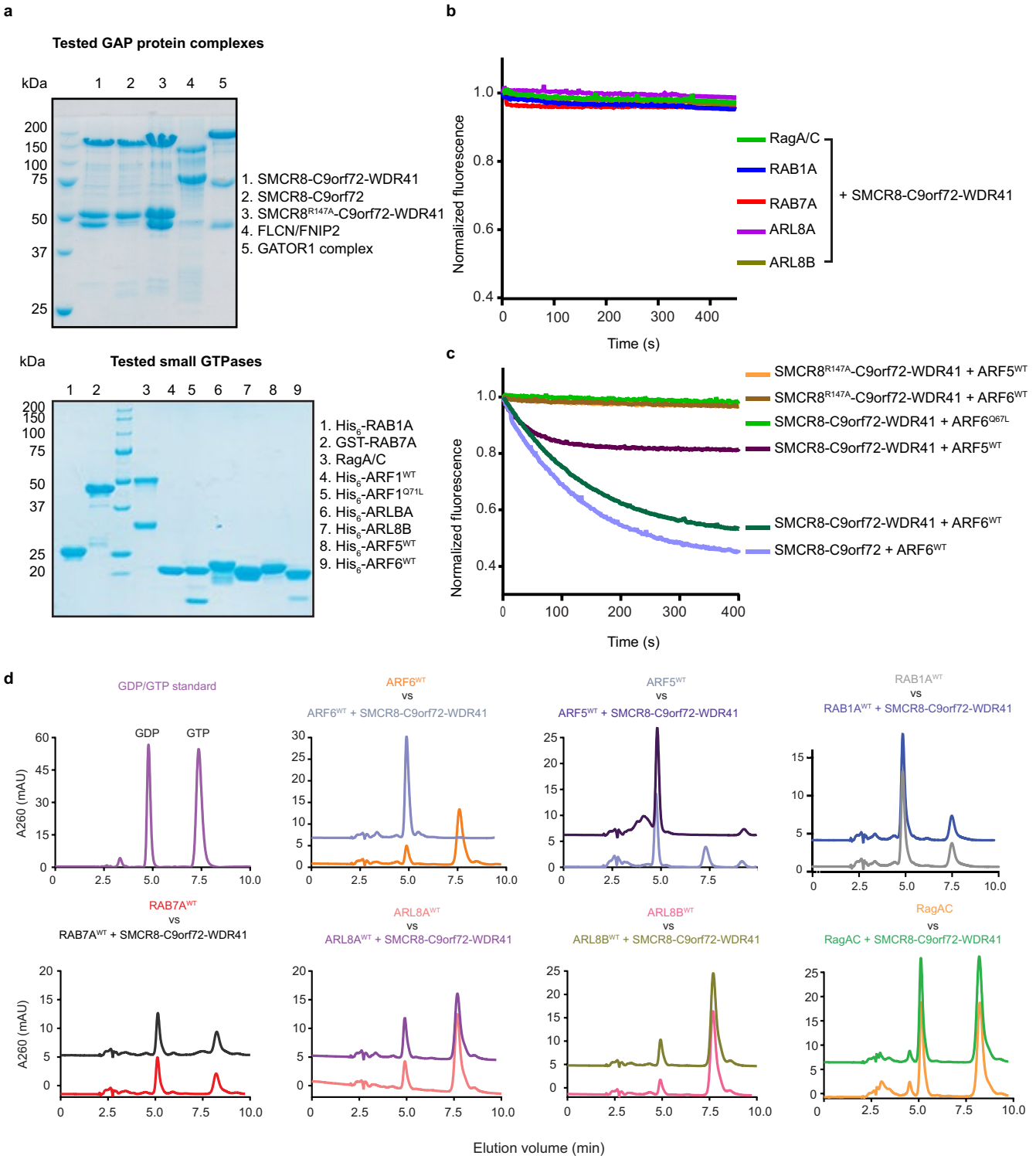
Extended Data Fig. 6 | Mapping the protected region from HDX results onto the SMCR8–C9orf72–WDR41 structure. a, The HDX uptake difference at 0.5 s was mapped on C9orf72–SMCR8. **b**–**d**, Close-up view of M1 (**b**) and M2–M5 (**c**) regions of SMCR8, and the M1 region of C9orf72 (**d**). **e**, Enlarged view of the WDR41 residues in the SMCR8–WDR41 interface.



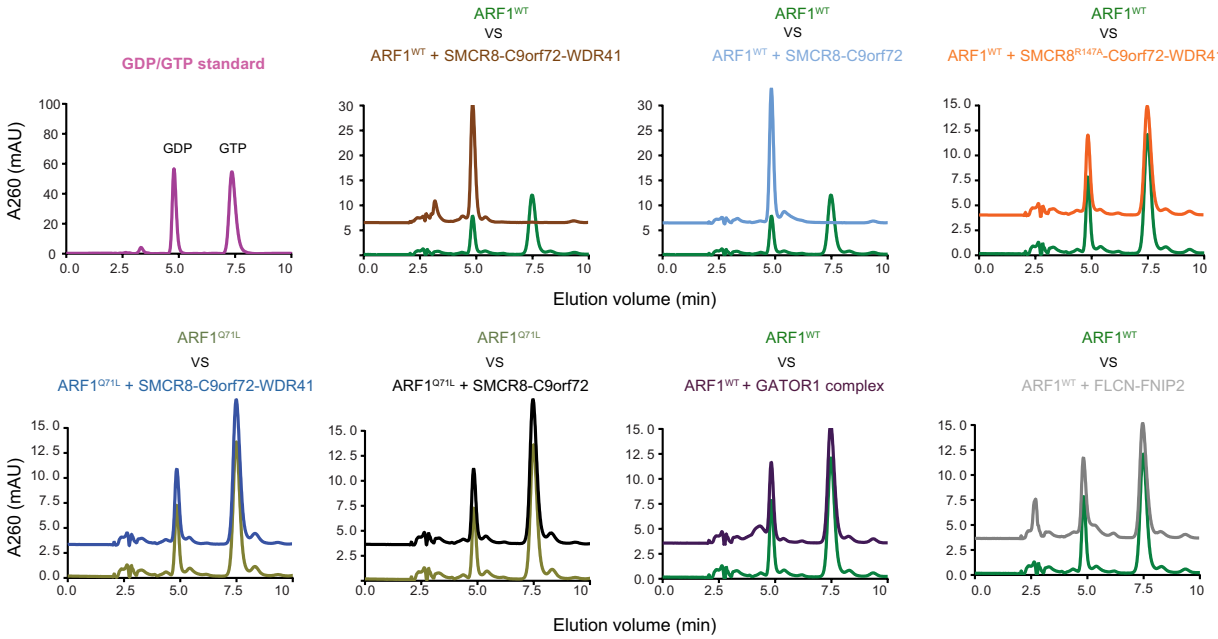
Extended Data Fig. 7 | Co-expression and pulldown validation of C9orf72–SMCR8 and SMCR8–WDR41 interface. a, Close-up view of the residues that mediate the DENN– DENN dimerization between C9orf72 and SMCR8. **b**, Co-expression and pulldown experiment of Strep-tagged SMCR8 with GST–C9orf72 mutants and WDR41. **c**, Pulldown experiment of GST–WDR41 mutants with C9orf72–SMCR8. The pulldown experiments were carried out at least twice with similar results (**b**, **c**).



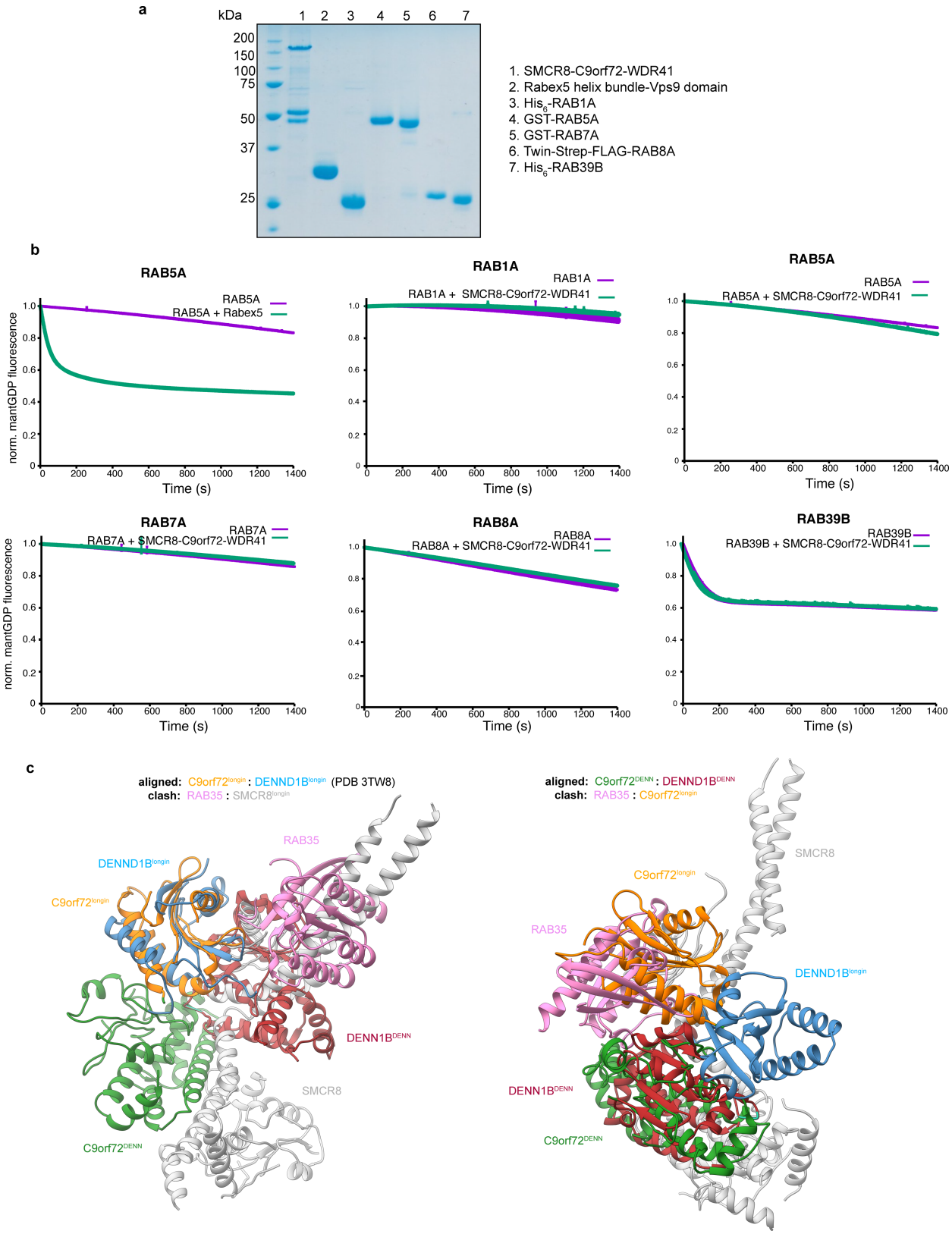
Extended Data Fig. 8 | Structural comparison between C9orf72–SMCR8 and FNIP2–FLCN. Left, structural alignment of full-length FNIP2–FLCN and C9orf72–SMCR8. Right, comparison between the longin dimers (top), DENN dimers (middle) and SMCR8 DENN with C9orf72 DENN domain (bottom).



Extended Data Fig. 9 | GTPase assay for different small GTPases with C9orf72–SMCR8–WDR41. a, SDS–PAGE of GAP protein complex (top) and GTPase proteins (bottom) used in the experiments. **b**, Tryptophan fluorescence GTPase signal was measured for purified RAGA–RAGC, ARL8A, ARL8B, RAB1A and RAB7A before and after addition of C9orf72–SMCR8–WDR41. The fluorescence signal upon GAP addition was normalized to 1 for each experiment. The experiments were carried out in triplicate and one representative plot is plotted. **c**, Tryptophan fluorescence GTPase signal was measured for purified ARF6, ARF(Q67L) or ARF5, before and after addition of C9orf72–SMCR8–WDR41, C9orf72–SMCR8(R147A)–WDR41 or C9orf72–SMCR8. **d**, HPLC-based GTPase assay with ARF6, ARF5, RAB1A, RAB7A, ARL8A, ARL8B and RAGA–RAGC proteins in the absence and addition of C9orf72–SMCR8–WDR41 complex, as indicated. The experiments were carried out in triplicate and one representative plot is shown. All experiments were carried out at least three times independently with similar results (**a**–**d**).



Extended Data Fig. 10 | HPLC-based GTPase assay with ARF1 or ARF1(Q71L) proteins. Assays were performed in the absence and addition of GAP complex, as indicated. The experiments were carried out in triplicate and one representative plot is shown. All experiments were carried out at least three times independently with similar results.



Extended Data Fig. 11 | GEF assay for different small GTPases with C9orf72–SMCR8–WDR41. a, SDS–PAGE of C9orf72–SMCR8–WDR41 complex and GTPase proteins used in the experiments. **b**, GEF assay with mantGDP-reloaded RAB1A, RAB5A, RAB7A, RAB8A and RAB39B proteins in the absence and addition of C9orf72–SMCR8–WDR41 complex, as indicated. RAB5A treated with RABEX5 was used as a positive-control reaction. Data were baseline-subtracted and normalized to the signal immediately after GTP addition, which also is the 0-s time point in the plots. Plots are mean ± s.d. of each technical triplicate experiment. All experiments were carried out at least twice independently with similar results (**a**, **b**). **c**, Structural alignment of C9orf72–SMCR8–WDR41 with DENND1B–RAB35 (PDB 3TW8).

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	C9orf72-SMCR8-WDR41 (EMDB-21048) (PDB 6V4U)
Data collection and processing	
Magnification (calibrated)	43,516
Voltage (kV)	300
Electron exposure (e–/Å^2)	59.6
Defocus range (μm)	0.06
Pixel size (Å)	1.149
Symmetry imposed	C1
Initial particle images (no.)	4,810,184
Final particle images (no.)	381,450
Map resolution (Å)	3.80
FSC threshold	0.143
Map resolution range (Å)	3.3-11
Refinement	
Initial model used (PDB code)	-
Model resolution (Å)	4.5
FSC threshold	0.5
Model resolution range (Å)	n.a.
Map sharpening *B* factor (Å^2)	-50
Model composition	
Non-hydrogen atoms	7,073
Protein residues	1,106
Ligands	0
B factors (Å^2)	
Protein	108.36
Ligand	
R.m.s. deviations	
Bond lengths (Å)	0.002
Bond angles (°)	0.472
Validation	
MolProbity score	1.60
Clashscore	4.14
Poor rotamers (%)	0
Ramachandran plot	
Favored (%)	93.89
Allowed (%)	6.11
Disallowed (%)	0.00

nature research

Corresponding author(s): James H. Hurley

Last updated by author(s): 2020/05/06

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM 3.6.14
Data analysis	For Cryo-EM data processing, the following software were used: Focus, MotionCor2, gctf-v1.06, gautomatch, Relion 3.0-beta, cryoSPARC, cryoSPARCV2, For model building, the following software were used: I-TASSER, Modeller, RaptorX, Phyre2, Coot 0.9 and 0.8.9.1, Phenix , PyMOL v1.7.2.1, UCSF Chimera 1.13 and 1.14, UCSF pyem, Molprobity (as part of Phenix) For HDX analysis, the following software were used: Orbitrap Discovery mass spectrometer, HDExaminer, Proteome Discoverer 2.1 For data display, the following software were used: Prism 6,7,8 and Origin 6.0 Lysosomal enrichment was scored using a home-built Matlab script.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

EM density map has been deposited in the EMDB with accession number EMD-21048. Atomic coordinates for the C9orf72-SMCR8-WDR41 have been deposited in the PDB with accession number 6V4U.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. The cryoEM data was collected in one session with more than 3000 movies. The high particle number included in the final reconstruction (381,450) indicates that based on similar methods and analysis that are widely published the sample size was not limiting.
Data exclusions	Electron microscopy: micrographs were screened manually and bad ones discarded. After CTF estimation, the bad movies which were bad focus or included ice contamination were excluded. HDX-MS: some peptides are with low signal therefore not included in analysis
Replication	HDX experiment was performed technically triplicate or quadruplicate for 0.5, 5, 50, 500 and 50,000 sec timepoints, . GAP and GEF assay were carried out either technical or biological triplicates. All replicates were successful and yielded similar results.
Randomization	Samples were not allocated into groups. Randomization is not relevant to this study.
Blinding	Blinding was not relevant to this study because there was no group allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK 293 Gn-Ti and HEK293A cell lines are purchased from UC Berkeley Cell Culture Facility.
Authentication	The cell lines were authenticated using short tandem repeat analysis.
Mycoplasma contamination	Both cell lines were tested negative for mycoplasma contamination using a nuclear stain by fluorescence microscopy.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

A prion-like domain in ELF3 functions as a thermosensor in *Arabidopsis*

https://doi.org/10.1038/s41586-020-2644-7

Received: 15 August 2019

Accepted: 15 July 2020

Published online: 26 August 2020

Check for updates

Jae-Hoon Jung[1,2,8], Antonio D. Barbosa[1,8], Stephanie Hutin[3,8], Janet R. Kumita[4,5], Mingjun Gao[1], Dorothee Derwort[1], Catarina S. Silva[3], Xuelei Lai[1,3], Elodie Pierre[3], Feng Geng[1], Sol-Bi Kim[2], Sujeong Baek[2], Chloe Zubieta[3], Katja E. Jaeger[1,6] & Philip A. Wigge[1,6,7] ✉

Temperature controls plant growth and development, and climate change has already altered the phenology of wild plants and crops[1]. However, the mechanisms by which plants sense temperature are not well understood. The evening complex is a major signalling hub and a core component of the plant circadian clock[2,3]. The evening complex acts as a temperature-responsive transcriptional repressor, providing rhythmicity and temperature responsiveness to growth through unknown mechanisms[2,4–6]. The evening complex consists of EARLY FLOWERING 3 (ELF3)[4,7], a large scaffold protein and key component of temperature sensing; ELF4, a small α-helical protein; and LUX ARRYTHMO (LUX), a DNA-binding protein required to recruit the evening complex to transcriptional targets. ELF3 contains a polyglutamine (polyQ) repeat[8–10], embedded within a predicted prion domain (PrD). Here we find that the length of the polyQ repeat correlates with thermal responsiveness. We show that ELF3 proteins in plants from hotter climates, with no detectable PrD, are active at high temperatures, and lack thermal responsiveness. The temperature sensitivity of ELF3 is also modulated by the levels of ELF4, indicating that ELF4 can stabilize the function of ELF3. In both *Arabidopsis* and a heterologous system, ELF3 fused with green fluorescent protein forms speckles within minutes in response to higher temperatures, in a PrD-dependent manner. A purified fragment encompassing the ELF3 PrD reversibly forms liquid droplets in response to increasing temperatures in vitro, indicating that these properties reflect a direct biophysical response conferred by the PrD. The ability of temperature to rapidly shift ELF3 between active and inactive states via phase transition represents a previously unknown thermosensory mechanism.

Arabidopsis ELF3 contains a polyQ repeat that varies in length from 7 to 29 residues in natural populations, and has previously been associated with phenotypic variation[8–10] (Fig. 1a). We therefore investigated whether the length of the polyQ repeat influences ELF3 activity. We found that in an isogenic Col-0 (wild-type) background, complementing *elf3-1* with *ELF3* transgenes that encode increasing polyQ lengths enhances thermoresponsiveness, as measured by hypocotyl elongation (Extended Data Figs. 1a, 2). However, the effects of altering polyQ length are mild, in agreement with other studies[8,11], and lines without a polyQ tract are still thermally responsive. This indicates that other features of ELF3 are also required to respond to temperature. As the polyQ repeat is located in the centre of a region that is predicted to be a prion domain[12] (PrD; residues 430–609) (Fig. 1a), we hypothesized that this domain might confer temperature responsiveness.

If the PrD of ELF3 does play a part in temperature sensing in *Arabidopsis*, we wondered whether this region varies in plants that are adapted to different climates. Indeed, ELF3 from *Solanum tuberosum*, which grows in temperate climates, has a smaller predicted PrD compared with *Arabidopsis*, whereas *Brachypodium distachyon*, which is habituated to warmer climates, is not predicted to have a PrD region (Fig. 1a and Extended Data Fig. 1b). As accelerated flowering is a major adaptive response of *Arabidopsis* to warm temperature, we investigated whether *Solanum tuberosum ELF3* (*StELF3*) and *Brachypodium distachyon ELF3* (*BdELF3*) alter this trait. *BdELF3* and *StELF3* are functional in *Arabidopsis* and complement *elf3-1* (Extended Data Fig. 3). At 22 °C, these plants resemble the wild type. However, at 27 °C they lose almost all of their thermally responsive early flowering (Fig. 1b); thus, these ELF3 variants with reduced or undetectable PrDs are largely unable to respond to warm temperatures. To test whether the thermal responsiveness of *Arabidopsis* ELF3 is due to the PrD itself, we created a chimeric version, in which we replaced the PrD of *Arabidopsis* with the corresponding sequence of BdELF3 (Extended Data Fig. 1b). Chimeric ELF3–BdPrD shows a suppression of temperature-responsive flowering, confirming that the PrD from *Arabidopsis* confers thermal responsiveness (Fig. 1b).

[1]Sainsbury Laboratory, University of Cambridge, Cambridge, UK. [2]Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea. [3]Laboratoire de Physiologie Cellulaire and Végétale, Université Grenoble Alpes/CNRS/CEA/INRAE, Grenoble, France. [4]Department of Chemistry, University of Cambridge, Cambridge, UK. [5]Department of Pharmacology, University of Cambridge, Cambridge, UK. [6]Leibniz-Institut für Gemüse- und Zierpflanzenbau, Großbeeren, Germany. [7]Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany. [8]These authors contributed equally: Jae-Hoon Jung, Antonio D. Barbosa, Stephanie Hutin. ✉e-mail: wigge@igzev.de

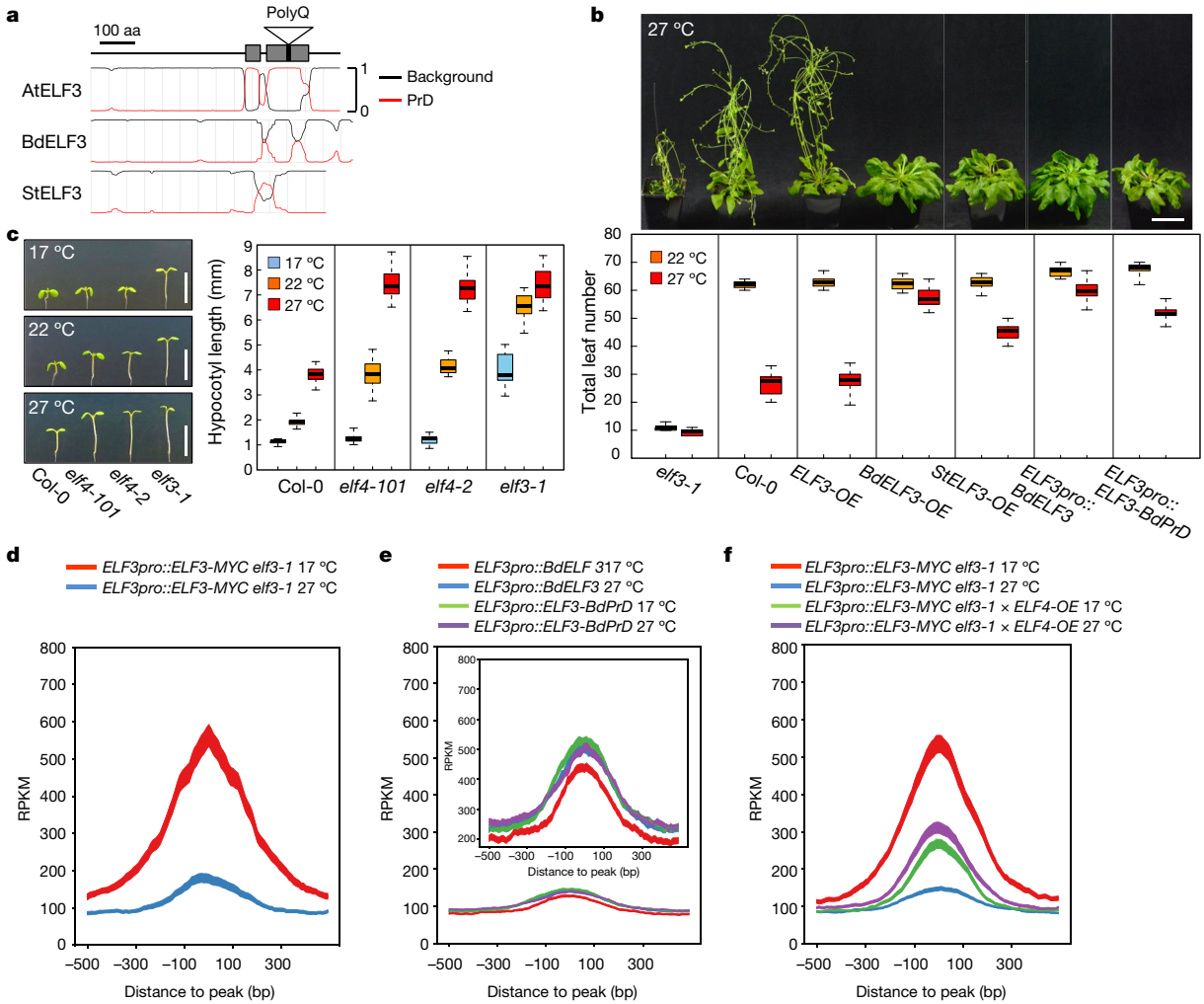


Fig. 1 | The polyQ repeat of ELF3 is embedded within a predicted prion domain that is essential for thermal responsiveness and modulated by ELF4. a, *Arabidopsis thaliana* ELF3 (AtELF3) contains a polyQ repeat embedded with a predicted prion domain (PrD, denoted by the red line reaching 1), whereas this PrD signature is absent (or below 1) in ELF3 from *Brachypodium distachyon* (BdELF3) and intermediate in *Solanum tuberosum* (StELF3). Prion domains were predicted using the Prion-Like Amino Acid Composition (PLAAC) algorithm[12]. **b**, The ELF3 PrD is essential to the thermal control of flowering time. Overexpressing AtELF3 (*ELF3-OE*) does not change the thermal induction of flowering, indicating that simply increasing the ELF3 protein level is not sufficient to delay flowering in response to higher temperatures. However, overexpressing BdELF3 causes an almost complete loss of thermal induction, and overexpressing StELF3 has a milder influence. These effects are also apparent when BdELF3 is expressed using the native ELF3 promoter (*ELF3pro::BdELF3*), and are dependent on the PrD, as replacing this domain of AtELF3 with the corresponding *B. distachyon* sequence (*ELF3pro::ELF3-BdPrD*) is sufficient to greatly reduce the thermal induction of flowering. **c**, At lower temperatures, *ELF4* becomes dispensable for controlling hypocotyl elongation. **d**, Binding of ELF3 to target genes, as measured by ChIP–seq, depends on temperature and declines genome-wide at 27 °C. **e**, Transgenic plants with stabilized forms of ELF3 do not respond to temperature. **f**, Overexpression of *ELF4* stabilizes ELF3 and removes the temperature responsiveness of its binding to targets. In box plots (**b**, **c**), each box is bounded by the lower and upper quartiles, the central bar represents the median, and the whiskers indicate minimum and maximum values. Scale bars, 5 mm (**b**, **c**).

The activity of ELF3 is modulated by binding to the small protein ELF4 (ref. [13]). To understand whether ELF4 contributes to the thermal responsiveness of ELF3, we investigated the effect of temperature on hypocotyl elongation and flowering time in *elf4-101* and *elf4-2* plants. At 22 °C and 27 °C, the effects of *elf4* alleles largely resemble those of *elf3-1* alleles, consistent with the key role for ELF4 in the evening complex[2,13,14]. At 17 °C, *ELF4* becomes dispensable for controlling both hypocotyl elongation and flowering time, and *elf4-101* and *elf4-2* plants have similar phenotypes to Col-0 plants (Fig. 1c and Extended Data Fig. 4a). These results suggest that ELF4 has a role in buffering the temperature responsiveness of ELF3 at higher temperatures, leading us to hypothesize that overexpressing *ELF4* may stabilize ELF3, as suggested by in vitro studies[15]. *ELF4* expression is regulated in a circadian manner, peaking at the end of the day and rapidly declining during the night[14]. We observed that plants constitutively overexpressing *ELF4* are largely unable to respond to temperature, as seen by both hypocotyl elongation and flowering time (Extended Data Fig. 4b), indicating that the presence of higher levels of ELF4 is sufficient to maintain ELF3 in an active state even at 27 °C (Extended Data Fig. 4b). This appears to be a consequence of modulating ELF3 function, as overexpressing *ELF4* has no detectable effect in the *elf3-1* background, and *ELF3* overexpression does not change thermal responsiveness (Extended Data Fig. 4c). ELF4 binds to a low-complexity region of ELF3 adjacent to the PrD (Extended Data Fig. 4d, e).

As ELF3 is a temperature-dependent transcriptional repressor, we sought to determine whether the phenotypic variation in responsiveness to temperature can be accounted for by variation in the occupancy of ELF3 on target genes. As shown previously[3,4], we found that the occupancy of ELF3 on target genes decreases as temperature increases (Fig. 1d and Extended Data Fig. 5). Consistent with our phenotypic

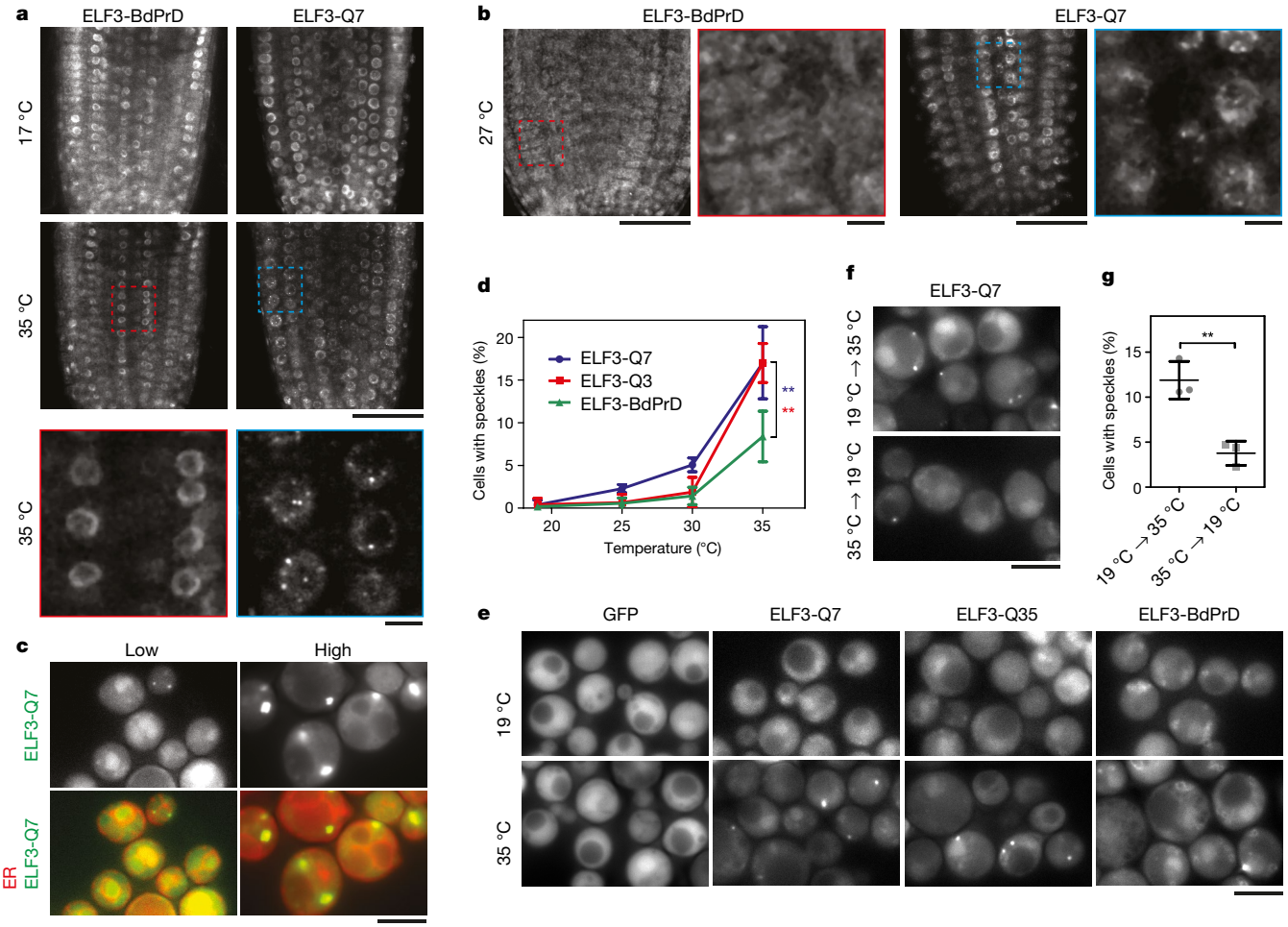


Fig. 2 | High temperature induces the formation of ELF3 speckles in vivo. **a**, Seedlings expressing GFP-tagged ELF3 with a seven-repeat polyQ sequence (ELF3-Q7) or a chimeric ELF3 in which the PrD was replaced by the corresponding region of *B. distachyon* ELF3 (ELF3-BdPrD) were grown in short photoperiods for 7 days at 17 °C. Roots were imaged by confocal microscopy before and after incubation at 35 °C for 15 min. $n = 4$, from two independent experiments. Scale bars in main images, 40 µm. Scale bars in magnified images, 5 µm. **b**, Seedlings as in **a**, but shifted to 27 °C for 2 h. $n = 4$ from two independent experiments. **c**, *Saccharomyces cerevisiae* cells expressing GFP-tagged ELF3 (Q7) from a centromeric plasmid ('Low') or an episomal plasmid ('High'), to achieve low or high expression levels of the protein, were grown in selective synthetic complete medium at 30 °C. Sec63-mCherry was used as an endoplasmic-reticulum (ER) reporter. Results representative of three independent experiments. Scale bars, 5 µm. **d**, Quantification of ELF3 speckles in *S. cerevisiae* cells expressing GFP fused to ELF3 (Q7), ELF3 with a longer polyQ repeat (Q35), or ELF3 with the PrD of *B. distachyon* (BdPrD), grown overnight at 19 °C and incubated at the indicated temperatures for 30 min. **e**, Representative images of cells from **d** at 19 °C and 35 °C. Scale bars, 5 µm. **f**, *Saccharomyces cerevisiae* cells expressing ELF3 were grown at 19 °C overnight and incubated at 35 °C for 30 min, and then shifted to 19 °C for 60 min. Cells shifted from 35 °C to 19 °C show a loss of speckles. Scale bars, 5 µm. **g**, Quantification of ELF3 speckles for the cells in **f**. **d**, **g**, Data shown as means ± s.d. $^{**}P < 0.01$. **d**, **e**, Results from two independent experiments for ELF3-Q7 at 19 °C, 25 °C and 30 °C; from three independent experiments for ELF3-Q35 and ELF3-BdPrD at 19 °C, 25 °C and 30 °C; and from four independent experiments (Q35), five independent experiments (Q7) or six independent experiments (BdPrD) at 35 °C. From these experiments, at least 100 cells were analysed for the presence of speckles. Statistical analyses were performed with Welch's two-sample *t*-test, with $P = 0.0062$ (Q7 versus BdPrD) and $P = 0.0016$ (Q35 versus BdPrD). **f**, **g**, Results from three independent experiments, counting at least 200 cells. Statistical analyses were performed with Welch's two-sample *t*-test with $P = 0.0075$.

observations, we also found that forms of ELF3 that lack a PrD also lose their temperature responsiveness of binding, indicating that the PrD directly modulates the thermoresponsiveness of ELF3 binding at target genes (Fig. 1e and Extended Data Fig. 5). Finally, overexpression of *ELF4* was also sufficient to stabilize ELF3 binding and abolish the temperature response (Fig. 1f and Extended Data Fig. 5).

As ELF3 functions as a transcriptional regulator, we sought to determine whether the observed changes in occupancy have a detectable influence on the *ELF3*-dependent transcriptome. We identified 325 *ELF3*-dependent marker genes by filtering for transcripts with a similar expression level to *LUX*, including key targets of the evening complex such as *PIF4* and *GI*. These genes are less thermally responsive in backgrounds expressing *BdELF3* or overexpressing *ELF4* (Extended Data Fig. 6a–c). We observed a mild effect of the polyQ lines on the expression of *ELF3*-dependent genes, consistent with the more subtle hypocotyl-length phenotypes of these lines (Extended Data Fig. 6a, b). We found 112 genes associated with the strongest ELF3 chromatin immunoprecipitation sequencing (ChIP–seq) peaks, and 25 of these are in common with the 325 *ELF3*-dependent genes, consistent with ELF3 dependence being a direct mechanism (Extended Data Fig. 6c and Supplementary Table). The *ELF3*-dependent genes directly bound by ELF3 show a clear temperature responsiveness in their expression, and this is affected when ELF3 binding is stabilized.

To investigate whether temperature controls ELF3 activity directly, we analysed the behaviour of natively expressed ELF3 fused to green fluorescent protein (GFP) in planta. At 17 °C, ELF3–GFP is nuclear with a diffuse signal. At 35 °C, multiple bright speckles form—behaviour that is specific to the presence of the PrD, as chimeric ELF3 with the

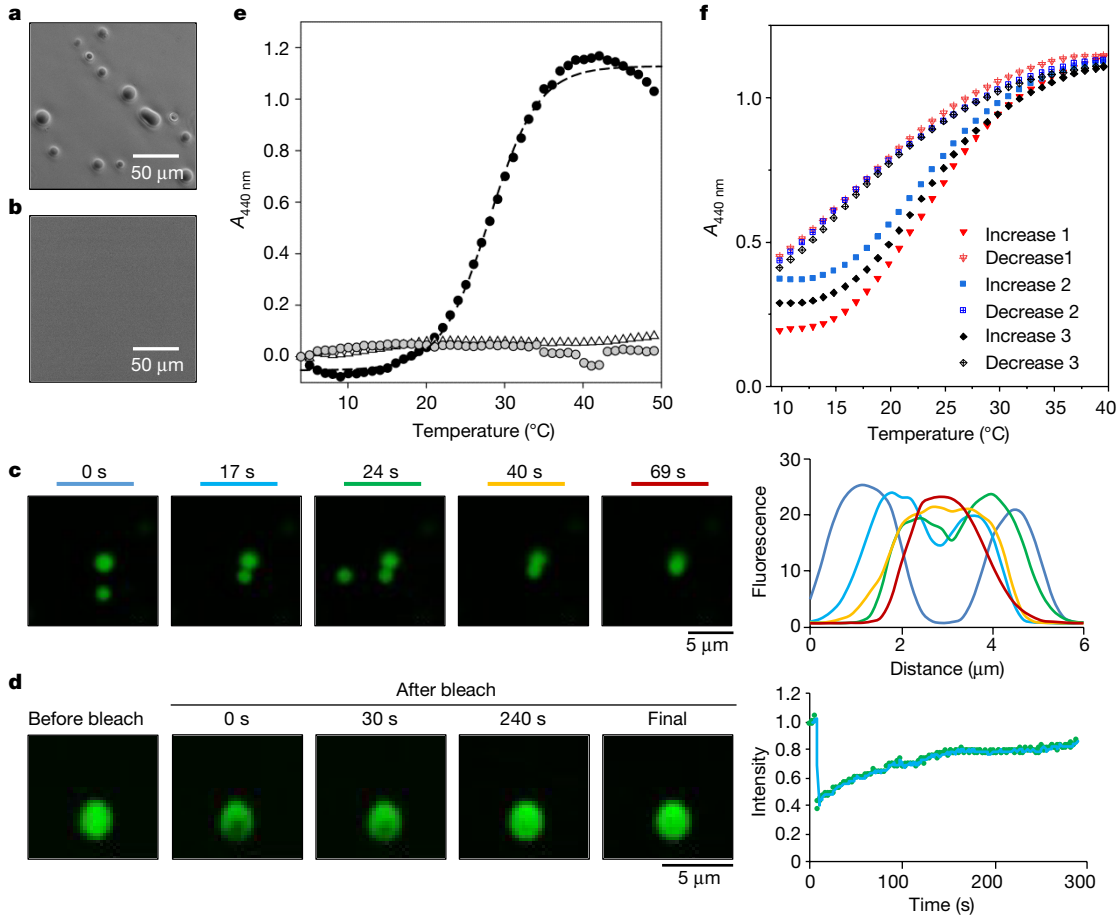


Fig. 3 | The PrD of ELF3 undergoes a reversible phase transition in response to temperature. a, Purified ELF3 PrD peptide forms liquid droplets in vitro. **b**, The equivalent protein domain from BdELF3, which is not predicted to contain a PrD, remains soluble and does not show any liquid droplet formation. **c**, Purified ELF3 PrD–GFP protein forms spherical droplets in vitro, which fuse. The graph shows how the signal intensity of initially discrete droplets occurs in two peaks, which merge into a single droplet over time. **d**, ELF3 PrD–GFP droplets show rapid recovery after photobleaching, indicating that they are liquid droplets. **e**, Light scattering assayed as a function of temperature for ELF3 PrD (15 µM; black circles), BdELF3 (15 µM; grey circles) and buffer alone (50 mM *N*-cyclohexyl-3-aminopropane sulfonic acid (CAPS) pH 9.7, 150 mM NaCl, 1 mM Tris-(2-carboxyethyl)phosphine (TCEP); open triangles). A_{440}, absorbance at 440 nm. The dashed line shows curve-fitting using a four-parameter sigmoidal equation. The transition temperature (T_m) for ELF3 PrD is 28.7 ± 1.8 °C, and the spectrum is representative of three independent experiments. **f**, Reversibility of light scattering as a function of temperature for ELF3 PrD (15 µM; in 50 mM CAPS pH 9.7, 200 mM NaCl, 1 mM TCEP). For one sample, the temperature was increased and decreased from 10 °C to 40 °C three times in succession (at the rate of 1 °C per min). The observed turbidity was reversible and consistently returned to A_{440} = 0.432 ± 0.02. The initial absorbance reading for repeat 3 (A_{440} = 0.288) is lower than for repeat 2 (A_{440} = 0.373) and this is probably because of time-dependent equilibration (as noted in **f**). The spectra are representative of two independent experiments and similar results were observed for 5 µM samples.

BdPrD remains largely diffuse in response to the warmer temperature (Fig. 2a). This behaviour is also observed at 27 °C (Fig. 2b). Increasing polyQ length also results in a greater tendency to form speckles (Extended Data Fig. 7).

As ELF3 in planta may be influenced by other factors that have co-evolved to control its activity, we sought to determine whether ELF3 expressed in *Saccharomyces cerevisiae*—a heterologous system lacking *ELF3*-related genes—is temperature-responsive. Under a low-expression system, ELF3–GFP forms a largely diffuse signal in yeast cells; when highly expressed, it forms bright puncta or speckles (Fig. 2c). We next investigated the influence of temperature on ELF3–GFP in yeast. At 19 °C, the signal is largely diffuse. Shifting cells to 35 °C results in a rapid formation of sharp punctate GFP signals; this phenomenon is more notable for ELF3 Q7 and Q35 proteins than for ELF3 BdPrD or free GFP (Fig. 2d, e). These effects appear to be specific to ELF3, as 35 °C is below the temperature required for endogenous proteins to aggregate[16], and we observed robust cell growth and protein expression under these conditions (Extended Data Fig. 8a, b). Although classically prions are associated with stable insoluble aggregates in the cell, the evening complex and ELF3 undergo diurnal cycles of activity, and temperatures fluctuate over short time scales, suggesting that reversibility of the temperature response is important. Indeed, the evening complex rapidly returns to full activity when plants are shifted from 27 °C to 22 °C (ref. [4]), leading us to hypothesize that the formation of speckles may be reversible. This is the case in yeast: returning cells from 35 °C to 19 °C results in a rapid reduction in the number of speckles (Fig. 2f, g).

These results are consistent with ELF3 being able to adopt two conformations: an active soluble form, and, at warmer temperatures, a higher-order multimeric form that is visible as bright speckles. It has been suggested that a major biological function of prion-like proteins is to act as environmental switches, because they are able to rapidly change conformation and form liquid droplets[17]. As ELF3 is largely insoluble in vitro, we investigated a peptide fragment spanning the PrD that we found to be soluble (ELF3 PrD; residues 388–625). *Arabidopsis thaliana* ELF3 PrD, in contrast to BdELF3 PrD, rapidly and reversibly forms liquid droplets as a function of ionic strength, protein concentration and temperature (Fig. 3a, b and Extended Data Fig. 9).

To analyse the dynamics of this behaviour, we purified ELF3 PrD fused to GFP (PrD–GFP). Decreasing the salt concentration and pH induces PrD–GFP to undergo a phase transition, forming micrometre-sized spherical droplets. The droplets are highly mobile in solution and are able to fuse, indicative of phase-separated liquids (Fig. 3c and Extended Data Fig. 10a, b). Using fluorescence recovery after photobleaching (FRAP), we found that recovery fractions ranged from 0.1 to 0.8, with recovery half-lives from seconds to minutes (Fig. 3d and Extended Data Fig. 10c, d). To determine whether the PrD is thermoresponsive, we analysed liquid droplet formation as a function of temperature. The purified ELF3 PrD peptide is more soluble at low temperatures, but undergoes a sharp phase transition with a midpoint at 28.7 ± 1.8 °C under our conditions (Fig. 3e). When the temperature is decreased, this process is largely reversible and the cycle can be repeated (Fig. 3f). This finding is in agreement with the temperature responsiveness of the ELF3 system in *Arabidopsis* seedlings, where the strongest phenotypic effects are observed at 27 °C. This response is specific for the PrD: the equivalent peptide fragment of BdELF3 shows no liquid droplet formation (Fig. 3e).

These results suggest that the PrD of ELF3 serves as a tunable thermosensor. Intrinsically disordered proteins frequently display thermoresponsive liquid–liquid phase separation—a behaviour driven by solvent-mediated interactions in a sequence-dependent manner[18]. As the PrD and intrinsically disordered protein sequences are widespread within eukaryotes[19], it will be interesting to see whether they have been recruited to provide thermosensory behaviour through phase transitions in other signalling contexts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2644-7.

1. Scheffers, B. R. et al. The broad footprint of climate change from genes to biomes to people. *Science* **354**, aaf7671 (2016).
2. Nusinow, D. A. et al. The ELF4–ELF3–LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature* **475**, 398–402 (2011).
3. Ezer, D. et al. The evening complex coordinates environmental and endogenous signals in *Arabidopsis*. *Nat. Plants* **3**, 17087 (2017).
4. Box, M. S. et al. ELF3 controls thermoresponsive growth in *Arabidopsis*. *Curr. Biol.* **25**, 194–199 (2015).
5. Mizuno, T. et al. Ambient temperature signal feeds into the circadian clock transcriptional circuitry through the EC night-time repressor in *Arabidopsis thaliana*. *Plant Cell Physiol.* **55**, 958–976 (2014).
6. Raschke, A. et al. Natural variants of ELF3 affect thermomorphogenesis by transcriptionally modulating PIF4-dependent auxin response genes. *BMC Plant Biol.* **15**, 197 (2015).
7. Nieto, C., López-Salmerón, V., Davière, J. M. & Prat, S. ELF3-PIF4 interaction regulates plant growth independently of the Evening Complex. *Curr. Biol.* **25**, 187–193 (2015).
8. Undurraga, S. F. et al. Background-dependent effects of polyglutamine variation in the *Arabidopsis thaliana* gene *ELF3*. *Proc. Natl Acad. Sci. USA* **109**, 19363–19367 (2012).
9. Tajima, T., Oda, A., Nakagawa, M., Kamada, H. & Mizoguchi, T. Natural variation of polyglutamine repeats of a circadian clock gene *ELF3* in *Arabidopsis*. *Plant Biotechnol.* **24**, 237–240 (2007).
10. Jiménez-Gómez, J. M., Wallace, A. D. & Maloof, J. N. Network analysis identifies ELF3 as a QTL for the shade avoidance response in *Arabidopsis*. *PLoS Genet.* **6**, e1001100 (2010).
11. Press, M. O. & Queitsch, C. Variability in a short tandem repeat mediates complex epistatic interactions in *Arabidopsis thaliana*. *Genetics* **205**, 455–464 (2017).
12. Lancaster, A. K., Nutter-Upham, A., Lindquist, S. & King, O. D. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**, 2501–2502 (2014).
13. Herrero, E. et al. EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the *Arabidopsis* circadian clock. *Plant Cell* **24**, 428–443 (2012).
14. Doyle, M. R. et al. The *ELF4* gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*. *Nature* **419**, 74–77 (2002).
15. Silva, C. S. et al. Molecular mechanisms of Evening Complex activity in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **117**, 6901–6909 (2020).
16. Wallace, E. W. J. et al. Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell* **162**, 1286–1298 (2015).
17. Franzmann, T. M. et al. Phase separation of a yeast prion protein promotes cellular fitness. *Science* **359**, eaao5654 (2018).
18. Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Temperature-controlled liquid–liquid phase separation of disordered proteins. *ACS Cent. Sci.* **5**, 821–830 (2019).
19. Si, K. Prions: what are they good for? *Annu. Rev. Cell Dev. Biol.* **31**, 149–169 (2015).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Generation of transgenic plants

Arabidopsis thaliana lines were in the Columbia (Col-0) background. The *elf3-1, elf4-2* and *elf4-101* mutants have previously been described[2–4]. To generate transgenic plants overexpressing *ELF4* (*ELF4-OE*), the *ELF4* coding sequence was subcloned into the pENTR-D-TOPO vector (ThermoFisher Scientific, Rockford, IL) according to the manufacturer's procedure. The resultant entry plasmid was recombined using LR clonase into the gateway binary pJHA212B vector, which contains the 35S promoter and carboxy-terminal 3× Flag tag sequences. The binary construct was transformed into Col-0 plants using the floral dipping method. The *ELF4-OE* transgenic plants were isolated by basta selection, and propagated to obtain single insertion lines with phenotypes of short hypocotyls and delayed flowering. The *ELF3-OE* transgenic plant has previously been described[4]. The *ELF3-OE* and *ELF4-OE* plants were crossed with *elf4-2* and *elf3-1* plants, respectively, and the resultant homozygous generations were used to measure hypocotyl length and flowering time.

To investigate whether the length of the polyQ repeat influences ELF3 activity, we first subcloned a 7.8-kilobase genomic fragment of *ELF3*, including its promoter and stop codon, into the pENTR-D-TOPO vector as above. The ELF3 protein in Col-0 plants has a polyQ repeat sequence of Q7. The Q7 repeat sequence in the entry plasmid was deleted or extended to Q21 using an overlapping polymerase chain reaction (PCR) strategy. Three kinds of entry plasmids encoding ELF3 proteins with Q0, Q7 or Q21 were recombined with LR clonase into the gateway binary pJHA212K vector without any tagging sequences. The binary construct was transformed into the *elf3-1* mutant through the floral dipping method. Three kinds of transgenic plants were isolated by kanamycin selection and propagated to obtain single insertion lines that rescue the long hypocotyl phenotype of *elf3-1* plants. The resultant homozygous generations were used to measure hypocotyl lengths.

To investigate whether the PrD in ELF3 confers thermal responsiveness, we generated transgenic plants expressing *StELF3* and *BdELF3* under the control of the native *A. thaliana ELF3* and 35S promoters in *elf3-1*. Genomic DNA was first isolated from the nuclei of *S. tuberosum* and *B. distachyon* using a standard cetyl trimethylammonium bromide (CTAB) DNA-extraction method. Coding sequences of *StELF3* and *BdELF3* were amplified by PCR using genomic DNA from *S. tuberosum* and *B. distachyon* as templates. The PCR fragments were cloned into SLIC binary vectors containing the 35S promoter and amino-terminal 3× Flag tag sequences using a NEBuilder HiFi DNA assembly cloning kit (New England BioLabs), and the constructs were transformed into the *elf3-1* mutant, resulting in *StELF3-OE* and *BdELF3-OE*, respectively. The same PCR fragments were also cloned into SLIC binary vectors containing the *A. thaliana ELF3* promoter and C-terminal 3× Flag tag sequences, and the constructs were transformed into the *elf3-1* mutant, resulting in *ELF3pro:StELF3* and *ELF3pro:BdELF3* transgenic plants. To create a chimeric version of *A. thaliana* ELF3, with the PrD replaced with the corresponding sequence from BdELF3, the existing entry plasmid containing the *A. thaliana ELF3* promoter and coding region was modified to replace the DNA fragment encoding the PrD sequence (residues 430–609) with the corresponding fragment of *BdELF3* (Extended Data Fig. 3). The modified entry plasmid was recombined using LR clonase into the gateway binary pJHA212K vector containing a C-terminal 3× Flag tag sequence, and the constructs were transformed into the *elf3-1* mutant, resulting in *ELF3pro:ELF3-BdPrD* transgenic plants. All transgenic plants containing DNA fragments from *S. tuberosum* or *B. distachyon* were isolated by kanamycin selection and propagated to obtain single insertion lines that rescued the long-hypocotyl phenotype of *elf3-1* mutants.

Different kinds of ELF3 entry plasmids, encoding *A. thaliana* ELF3 proteins with varying polyQ lengths (Q0–Q35) or *B. distachyon* ELF3 protein, were recombined into gateway binary pJHA212K vectors containing C-terminal 3× Flag or GFP tag sequences. The resulting constructs were transformed into the *elf3-1* mutant to generate transgenic plants used for ChIP–seq or plant fluorescence microscopy experiments. The *ELF3pro:ELF3-MYC elf3-1* transgenic plant used for ChIP–seq has previously been described[3].

Plant growth conditions

Arabidopsis seeds were sterilized and sown on standard half-strength Murashige and Skoog agar (MS-agar) plates at pH 5.7. Sterilized seeds were stratified for 3 days at 4 °C in the dark, and allowed to germinate for 24 h at 22 °C under cool-white fluorescent light at 170 μmol m^{-2} s^{-1}. The plates were then transferred to short-photoperiod conditions (8-h light and 16-h dark) at different temperatures for assays. For hypocotyl-length measurements, 7-day-old or 8-day-old seedlings—grown under short-photoperiod conditions at a light intensity of 80 μmol m^{-2} s^{-1}—were photographed and analysed using ImageJ software (http://rsbweb.nih.gov/ij/).

To measure flowering time, plants were grown in soil under short-photoperiod conditions at either 22 °C or 27 °C until flowering. Flowering times were determined by counting the numbers of rosette and cauline leaves at bolting. Twenty to thirty plants were counted and averaged for each measurement.

ChIP–seq experiments

Seedlings were grown for 10 days under short-photoperiod conditions at either 17 °C or 22 °C, and shifted to 27 °C for 2 h at Zeitgeber time ZT8. Next, 3 g of seedlings from each treatment were fixed under vacuum for 20 min in 1× phosphate-buffered saline (PBS; 10 mM PO_4^{3-}, 137 mM NaCl and 2.7 mM KCl) containing 1% formaldehyde (Sigma, catalogue number F8775)). The reaction was quenched by adding glycine to a final concentration of 62 mM. ChIP experiments were performed as described[20]. Anti-Myc agarose affinity gel antibody (Sigma, A7470), anti-haemagglutinin agarose (Sigma, A2095) or anti-Flag M2 affinity gel (Sigma, A2220) was used for immunoprecipitation (100 μl of resin suspension, equivalent to 50 μl of settled resin, was used for each ChIP). Sequencing libraries were prepared using a TruSeq ChIP sample preparation kit (Illumina, IP-202-1024) or using a NEBNext Ultra II DNA library prep kit (New England BioLabs) and samples were sequenced using an Illumina NextSeq 500 platform.

RNA-sequencing experiments

Seedlings were grown on plates for 10 days and collected at the indicated time points; 70 mg of seedlings were pooled per tube and total RNA was extracted using a MagMAX-96 total RNA isolation kit (ThermoFisher) according to the manufacturer's instructions. Libraries were prepared using a Lexogen QuantSeq 3′ mRNA-seq library prep FWD kit (Illumina) according to the manufacturer's instructions. The libraries were sequenced on an Illumina NextSeq 500 platform.

ChIP–seq and RNA-seq bioinformatic analysis

The following pipeline quantifies gene expression and ChIP–seq binding. For the pile-up figure (Fig. 1d–f), coverage values were extracted from reads per kilobase of transcript per million mapped reads (RPKM) bigwig outputs from the pipeline for the file 'chipseq_differential_binding.peak_list.csv'. Figure 1 is shaded with standard errors computed for each *x*-value.

Two ELF3 ChIP–seq libraries were compared with a shortlist of 362 1-base-pair genomic intervals that show reduced binding at 27 °C compared with 17 °C. These genomic intervals were used for pileup of other ChIP–seq libraries, generating file 'chipseq_targets_genes_job.peak_list.csv'.

Continuing from 'chipseq_differential_binding.peak_list.csv', genomic intervals were filtered out if they lacked an annotated start codon within 500 bp. These genes were then deposited into file 'chipseq_differential_binding.peak_list.csv'. For column 'chipseq_targets_genes_job', see 'chipseq_targets_genes_job.peak_list.csv', column 'signature_targets'.

A signature score was computed for each of roughly 36,000 annotated genes, according to the similarity of their expression levels to that of a signature gene, *LUX*, within 10 selected datasets. The top one per cent of genes were then selected to be 'signature_targets'. The signature score (*s*) is defined as:

s = <meanNorm(expression(target gene)),meanNorm(expression(*LUX*))>*s* = <meanNorm(expression(target gene)),meanNorm(expression(*LUX*))>

in which <expr1,expr2> <expr1,expr2> is the dot product taken over the selected datasets.

Gene-expression analysis by quantitative PCR

Gene transcript levels were determined by reverse transcription with quantitative PCR (RT–qPCR). Isolation of total RNA from appropriate plant materials was carried out using Trizol reagent (Thermo Fisher Scientific) according to the manufacturer's recommendations. First-strand complementary DNA was synthesized from 1.5 µg of total RNA using a RevertAid first-strand cDNA synthesis kit (Thermo Fisher Scientific) according to the manufacturer's recommendations. RT–qPCR reactions were performed in 96-well blocks with the QuantStudio 1 real-time PCR system (Thermo Fisher Scientific) using TOPreal qPCR 2× PreMIX (SYBR Green with high carboxyrhodamine (ROX), Enzynomics) in a final volume of 20 µl. PCR primer sets of *ELF3* forward (5′-TCTAGTCAGCCTTGTGGTGTG-3′) and *ELF3* reverse (5′-TCCTCTGATCATGCTGTGCC-3′), *FLOWERING LOCUS T* (*FT*) forward (5′-GGTGGAGAAGACCTCAGGAA-3′) and *FT* reverse (5′-GGTTGCTAGGACTTGGAACATC-3′), and *EUKARYOTIC TRANSLATION INITIATION FACTOR 4A1* (*EIF4A1*) forward (5′-TGACCACACAGTCTCTGCAA-3′) and *eIF4A1* reverse (5′-ACCAGGGAGACTTGTTGGAC-3′) were used for RT–qPCR of *ELF3*, *FT* and *EIF4A* genes. The values for each set of primers were normalized relative to *EIF4A1* (At3g13920). All RT–qPCR reactions were performed in three technical replicates using total RNA samples extracted from three independent biological replicates.

Plant fluorescence microscopy

Seeds were sown on MS-agar plates and stratified for two to three days at 4 °C in the dark. The plates were then transferred into short-photoperiod conditions and grown for seven days at 17 °C. Roots were imaged before and after incubation of the slides at 35 °C or 30 °C for 15 min, or after 2 h of incubation of the seedlings on prewarmed MS-agar plates at 27 °C, on a Zeiss LSM880 upright confocal microscope with a 20 × dry Plan-Apochromatic 0.8 NA objective lens, and acquired using ZEN 2.3 software (Carl Zeiss). GFP fluorescence was excited with a 488-nm line from an argon laser. Images were saved as .czi files and then imported to ImageJ software.

To calculate speckle scores based on fluorescence intensity (Extended Data Fig. 10), we selected root regions corresponding to the size of individual cells, and measured mean, standard deviation (s.d.) and maximum grey values in ImageJ. We assumed that speckle formation would lead to higher grey values, and that a higher frequency of speckles within the analysis region would increase the standard deviation. A speckle score was obtained by calculating the ratio of the maximum and mean grey values, normalized to the average of the mean grey values for all the regions measured in each root (to account for local intensity variation), and finally multiplied by the standard deviation:

Speckle score = [maximum (grey value)/mean(grey value)]/average (mean grey value for all regions in the root) × s.d.

Yeast fluorescence microscopy

Yeast cells (RS453 *MATα ade2-1 his3-11, 15 ura3-52 leu2-3112 trp1-1, URA3::YIplac211–SEC63–mCherry*)[21] were transformed with the plasmids indicated below using the lithium acetate method, and grown in synthetic defined medium containing 0.17% yeast nitrogen base (MP Biomedicals), 0.5% ammonium sulfate (Fisher Scientific), –Leu/–Trp DO supplement (Clontech), and 60 mg l^{-1} leucine or 40 mg l^{-1} tryptophan (Sigma), for plasmid selection. Cells were grown overnight at 19 °C, incubated at 25 °C, 30 °C or 35 °C for 30 min and, where indicated, reincubated at 19 °C for 60 min. Cells were imaged live in a Zeiss AxioImager. Z2 epifluorescence upright microscope with a 100× Plan-Apochromatic 1.4 NA objective lens (Carl Zeiss). Images were recorded using a large-chip sCMOS monocamera for sensitive fluorescence imaging (ORCA Flash 4.0v2, Hamamatsu), saved using Zeiss ZEN2.3 software (Blue edition, Carl Zeiss) and exported to ImageJ. Plasmids were as follows: *YCplac111–NOP–GFP*, with *GFP* under the control of the *NOP1* promoter in the *LEU2/CEN* vector; *YCplac111–NOP–cELF3-Q7–GFP*, a C-terminally GFP-tagged *ELF3* cDNA with a seven-glutatmine polyQ repeat under the control of the *NOP1* promoter in the *LEU2/CEN* vector; *YCplac111–NOP–cELF3-Q35–GFP*, a C-terminally GFP-tagged *ELF3* cDNA with a 35-glutamine polyQ repeat under the control of the *NOP1* promoter in the *LEU2/CEN* vector; *YCplac111–NOP–cELF3–BdPrD–GFP*, a C-terminally GFP-tagged *ELF3* cDNA with the PrD domain of *B. distachyon* under the control of the *NOP1* promoter in the *LEU2/CEN* vector; and *pGBKT7–cELF3-Q7–GFP*, a C-terminally GFP-tagged wild-type *ELF3* cDNA under the control of the *ADH1* promoter in the *TRP1/2µ* vector.

Yeast two-hybrid assays

Yeast two-hybrid assays were performed using the BD Matchmaker system (Clontech). The pGADT7 vector was used for the GAL4 activation domain, and the pGBKT7 vector was used for the GAL4 DNA-binding domain. Clontech's Y2H Gold yeast strain was used for transformation. ELF4 and ELF3 cDNA sequences were subcloned into pGBKT7 and pGADT7 vectors, respectively. Transformation of vector constructs into Y2H Gold cells was performed according to the manufacturer's instructions. Colonies obtained were streaked on selective medium without leucine, tryptophan, histidine and adenine (–LWHA).

ELF3 PrD constructs

Arabidopsis thaliana ELF3 PrD (residues 388–625, At2g25930) and *B. distachyon* ELF3 PrD (residues 432–669, BRADI_2g14290) were cloned into the expression vector pESPRIT2 (refs. [22,23]) using the *Aat*II and *Not*I sites. The plasmid contains an N-terminal 6× histidine tag followed by a TEV protease cleavage site. All proteins were overproduced in *Escherichia coli* strain BL21 Rosetta 2 (Novagen).

Protein expression and purification

BdELF3 PrD, AtELF3 PrD and AtELF3 PrD–GFP were expressed in *E. coli* strain BL21, which was induced with 1 mM isopropyl-β-d-thiogalactopyranoside (IPTG) at 18 °C overnight. Bacterial pellets were resuspended in resuspension buffer (100 mM CAPS pH 9.7, 300 mM NaCl, 30 mM imidazole, 1 mM TCEP; Sigma) plus complete protease-inhibitor cocktail (Roche). Cells were lysed by sonication and the lysates were centrifuged at 50,000*g* for 30 min at 4 °C. For AtELF3 PrD and AtELF3 PrD–GFP, the supernatants were applied to a Ni-NTA column. The bound proteins were washed with 20 column bed volumes (CV) of resuspension buffer and then with 20 CV of a high salt buffer (100 mM CAPS pH 9.7, 1 M NaCl, 30 mM imidazole and 1 mM TCEP) and eluted with 5 CV of elution buffer (100 mM CAPS pH 9.7, 300 mM NaCl, 300 mM imidazole and 1 mM TCEP). The fractions of interest were pooled and dialysed overnight at 4 °C in 50 mM CAPS pH 9.7, 400 mM NaCl and 1 mM TCEP. For BdELF3 PrD, the pellet was solubilized in 8 M urea, 100 mM CAPS pH 9.7 and 300 mM NaCl. A second centrifugation was carried out and the supernatant was applied to a Ni-NTA column

pre-equilibrated with equilibration buffer (8 M urea, 100 mM CAPS pH 9.7, 300 mM NaCl, 30 mM imidazole and 1 mM TCEP). The bound protein was washed with 20 CV of equilibration buffer and then with 20 CV of high-salt buffer for on-column refolding. The protein was eluted with 5 CV of elution buffer. Fractions of interest were pooled and dialysed overnight at 4 °C in 50 mM CAPS pH 9.7, 300 mM NaCl and 1 mM TCEP. Protein purity was determined via SDS–PAGE.

Liquid droplet formation

For formation of liquid droplets, the NaCl concentration of the dialysis buffer (50 mM CAPS pH 9.7, 400 mM NaCl and 1 mM TCEP) was gradually decreased using a step gradient at 4 °C. Droplets were visualized after dialysis. Images were acquired using a 20× objective (LUCPLFLN20 × PH1/0.45) on an epifluorescence inverted microscope (CKX41 model) equipped with a pE-300 Cool-LED camera.

AtELF3PrD–GFP FRAP

For droplet visualization and photobleaching experiments using AtELF3PrD–GFP protein, liquid droplet formation was induced by mixing 5 µl of 5 mg ml^{-1} AtELF3PrD–GFP with 5 µl 20 mM Tris pH 7.5, 100 mM NaCl and 1 mM TCEP on a glass slide. The drop was covered with a coverslip and quickly mounted onto an EclipseTi-E Nikon inverted microscope as part of the confocal spinning disk system, with a CSUX1-A1 Yokogawa confocal head, an Evolve EMCCD camera (Roper Scientific, Princeton Instruments) and a Nikon CFI Plan-APO VC 60×, 1.4 NA, oil-immersion objective, controlled using MetaMorph (Universal Imaging) software with the autofocus function enabled. For photobleaching experiments, droplets were allowed to adhere to the coverslip before photobleaching to minimize droplet movement during the experiment. Acquisition times were approximately 1 s per image. Droplet size was roughly 2–5 µm, with a bleaching area of roughly 1 µm for partial bleaching and approximately 5 µm for full bleaching. Time-lapse images were acquired at 530 nM. Droplet intensity profiles were measured manually for quantification of droplet fusion in ImageJ[24]. For FRAP experiments, regions of interest (bleached, unbleached and background) were selected in ImageJ and processed using the easyFRAP webtool[25]. Corrected intensities were fit to a single exponential curve in ImageJ.

Light-scattering assay

Light scattering was assessed using a Cary 100 UV-vis spectrometer (Agilent Technologies). The absorbance at 440 nm was monitored for samples containing buffer alone (50 mM CAPS pH 9.7, 150 mM NaCl, 1 mM TCEP), ELF3 PrD (15 µM) or BdELF3 (15 µM) in quartz cuvettes (path length 10 mm) with increasing temperatures (4–50 °C; 1 °C min^{-1}), and the spectra were normalized with respect to the ELF3 PrD. A transition temperature (T_m) was determined by fitting the spectrum with a four-parameter sigmoidal equation (Sigmaplot 11, Systat Software). Reported values are an average from three separate experiments. To assess reversibility, we monitored the turbidity while increasing the temperature (from 10 °C to 40 °C; 1 °C min^{-1}) and then decreasing the temperature (from 40 °C to 10 °C; 1 °C min^{-1}); this cycle was repeated three times in total (using 5 µM or 15 µM ELF3 PrD, 50 mM CAPS pH 9.7, 200 mM NaCl and 1 mM TCEP).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data for gene-expression analysis (RNA-seq) and protein–DNA interactions (ChIP–seq) have been deposited in the publicly available Gene Expression Omnibus (GEO) under accession code GSE137264 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137264). The raw data used in this study are available at https://osf.io/fn5um/.

Code availability

The code to produce the figures (Fig. 1d–f and Extended Data Figs. 5, 7–9) from the processed files is available at https://github.com/shouldsee/polyq-figures. To enable easier browsing, a static site is hosted at https://shouldsee.github.io/polyq-figures. The inhouse pipeline for mapping is available at https://github.com/shouldsee/synoBio.

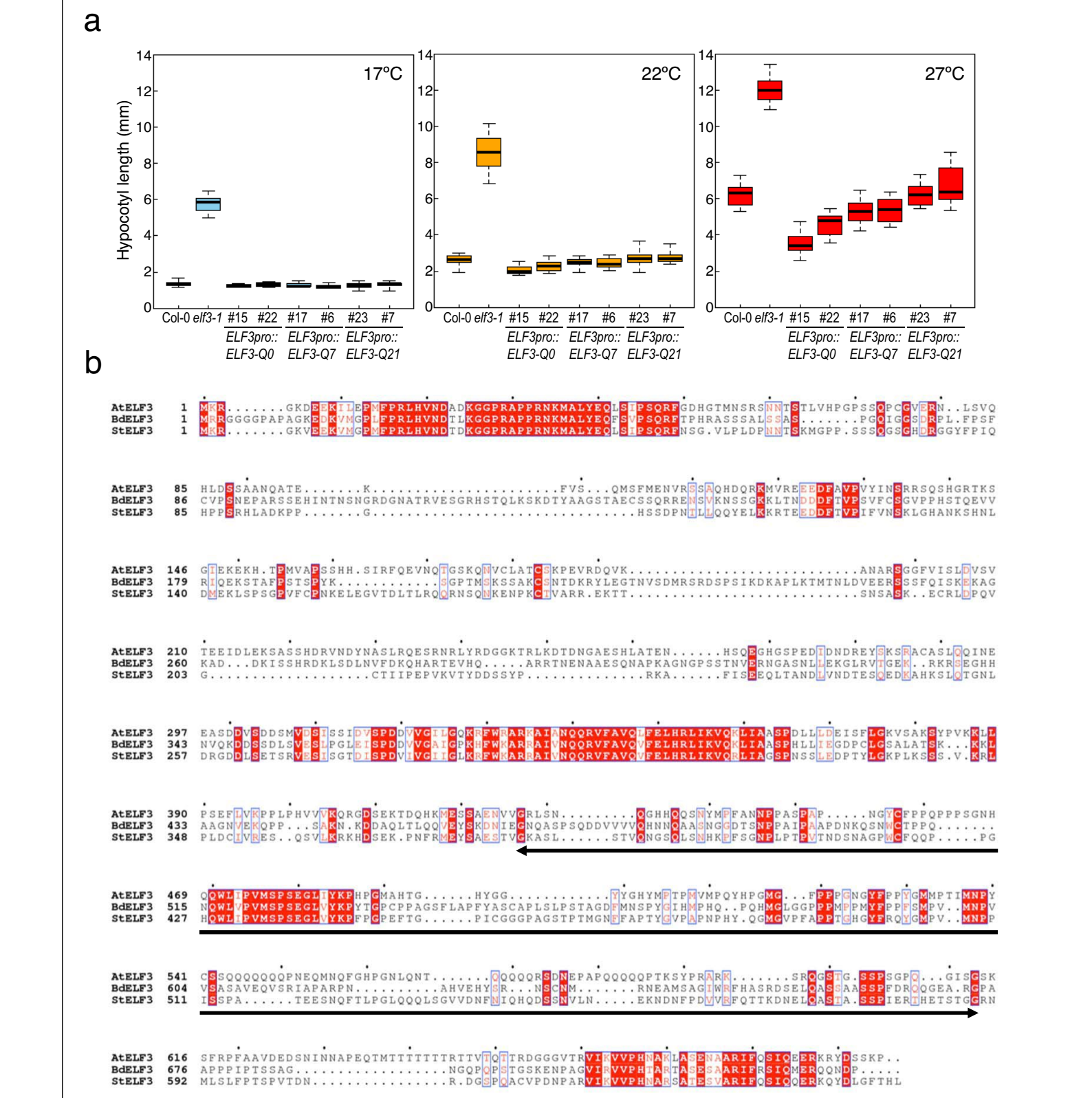
20. Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J. & Wigge, P. A. Interlocking feedback loops govern the dynamic behavior of the floral transition in *Arabidopsis*. *Plant Cell* **25**, 820–833 (2013).
21. Barbosa, A. D. et al. Lipid partitioning at the nuclear envelope controls membrane biogenesis. *Mol. Biol. Cell* **26**, 3641–3657 (2015).
22. Guilligay, D. et al. The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat. Struct. Mol. Biol.* **15**, 500–506 (2008).
23. Tarendeau, F. et al. Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat. Struct. Mol. Biol.* **14**, 229–233 (2007).
24. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
25. Koulouras, G. et al. EasyFRAP-web: a web-based tool for the analysis of fluorescence recovery after photobleaching data. *Nucleic Acids Res.* **46** (W1), W467–W472 (2018).

Acknowledgements We thank M. Perutz for discussions on polyQ proteins; the microscopy facility MuLife of the Interdisciplinary Research Institute of Grenoble (IRIG)/Department of Structural and Cellular Integrative Biology (DBSCI), funded by CEA Nanobio and Labex GRAL, for equipment access and use; and L. Kurzawa and F. Senger for technical assistance and discussions. This work used the platforms of the Grenoble Instruct-ERIC Center (ISBG; grant UMS 3518; CNRS/CEA/UGA/EMBL), with support from the French Infrastructure for Integrated Structural Biology (FRISBI; grant ANR-10-INBS-05-02) and Labex GRAL (grant ANR-CBH-EUR-GS) within the Grenoble Partnership for Structural Biology (PSB). P.A.W. and C.Z. received funding from Instruct-ERIC (PID 2236). This work was also supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (Ministry of Science and ICT; MSIT) (grant 2019R1C1C1010507 to J.-H.J). This work was further funded by project ANR-19-CE20-0021-01 PRC (to C.Z.). P.A.W. received support from the European Research Council (grant EC FP7 ERC 243140) and the Gatsby Charitable Foundation (grant GAT3273/GLB). P.A.W. receives funding from the Leibniz Foundation.

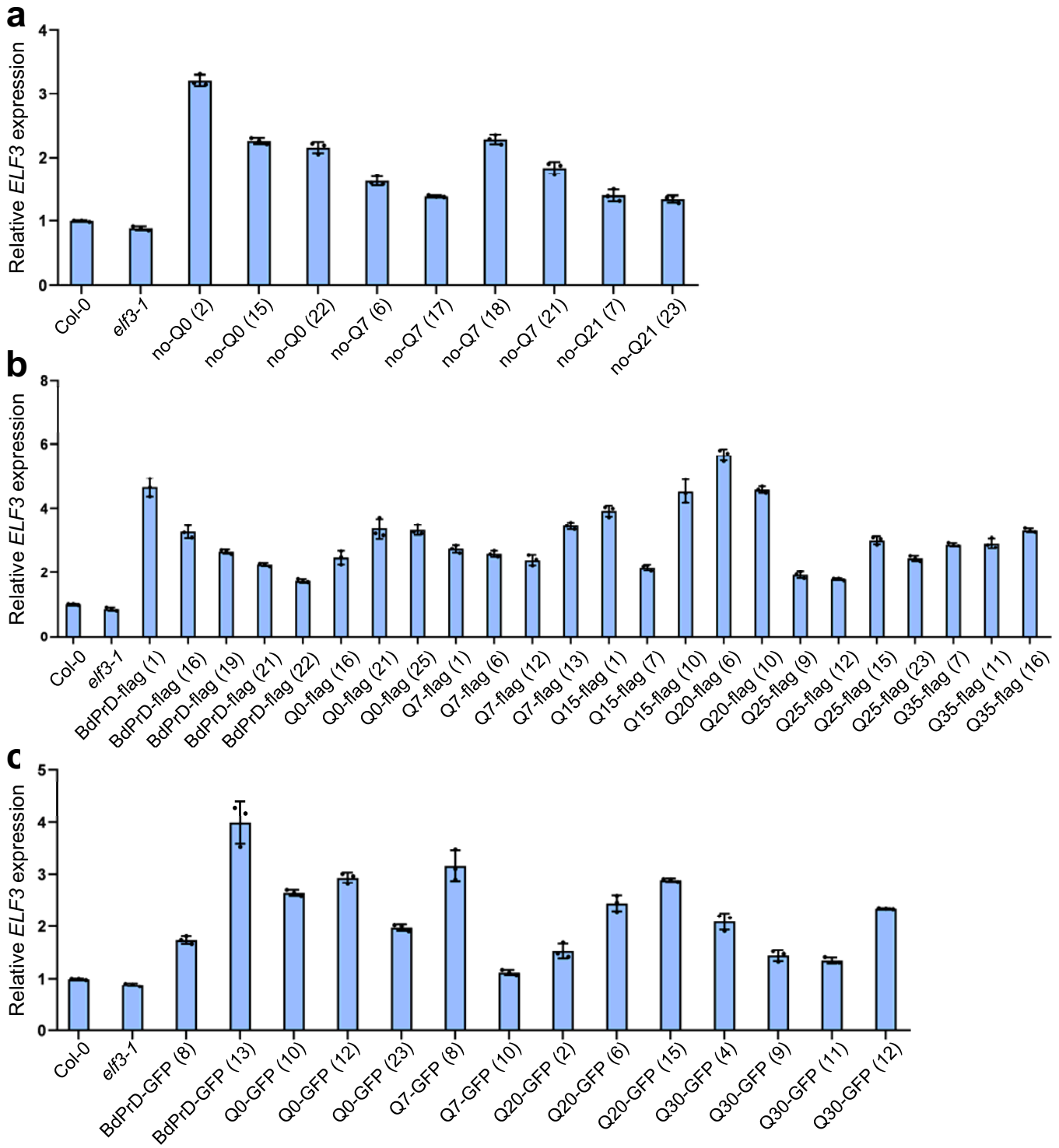
Author contributions J.-H.J., A.D.B., S.H., J.R.K., C.Z., K.E.J. and P.A.W. conceived the study and wrote the manuscript. J.-H.J. and M.G. generated transgenic plants and mutants and analysed their phenotypes. A.D.B. performed in vivo imaging experiments in yeasts and plants. S.H., J.R.K., C.S.S., X.L., E.P. and C.Z. performed and analysed in vitro phase-separation assays. K.E.J. performed ChIP–seq and RNA-seq experiments. D.D. and F.G. analysed sequencing data. S.-B.K. and S.B. performed yeast two-hybrid assays and analysed gene-expression levels in transgenic plants. C.Z., K.E.J. and P.A.W. supervised experimental work.

Competing interests The authors declare no competing interests.

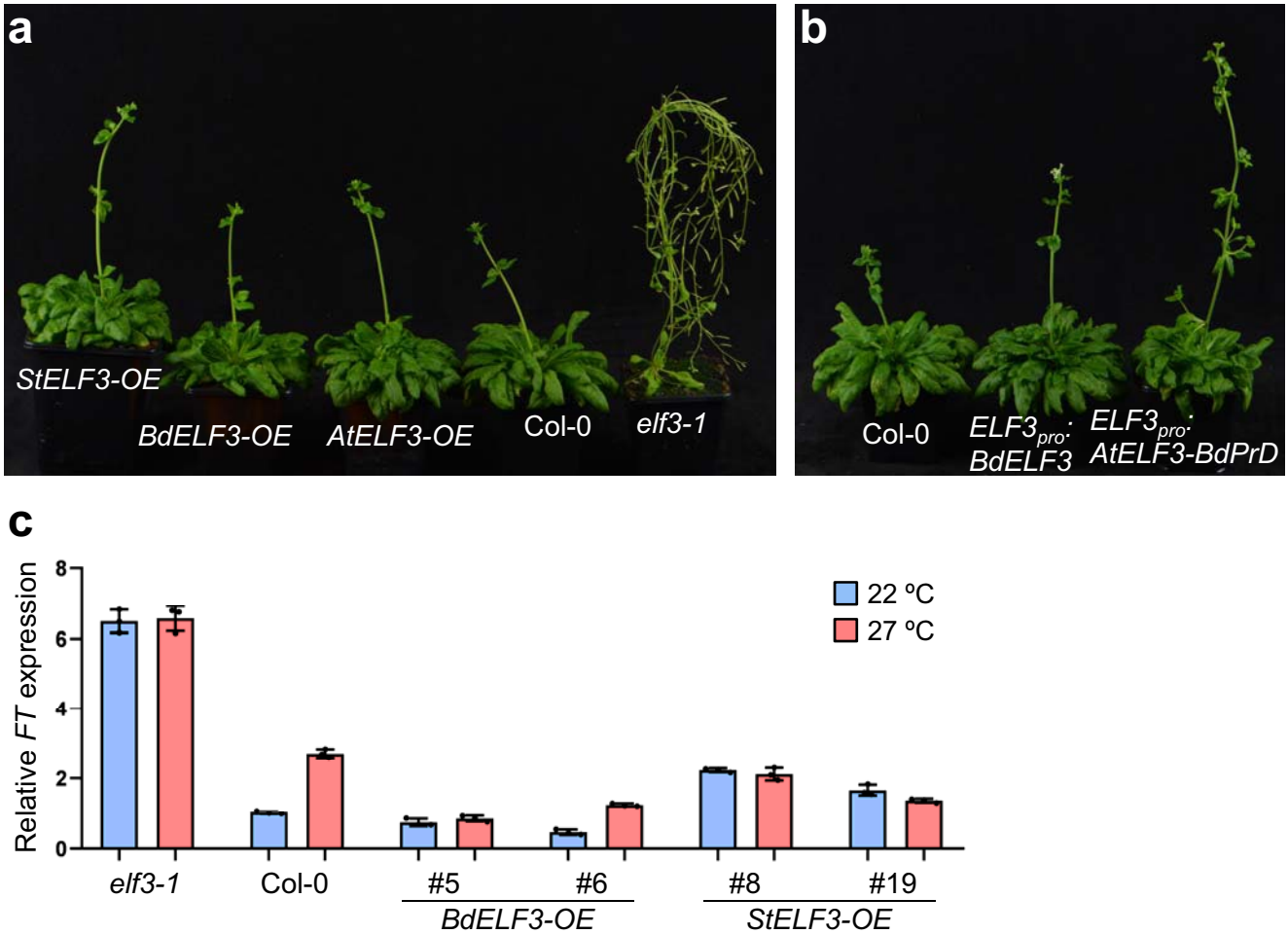
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2644-7.
Correspondence and requests for materials should be addressed to P.A.W.
Peer review information *Nature* thanks Salomé Prat and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



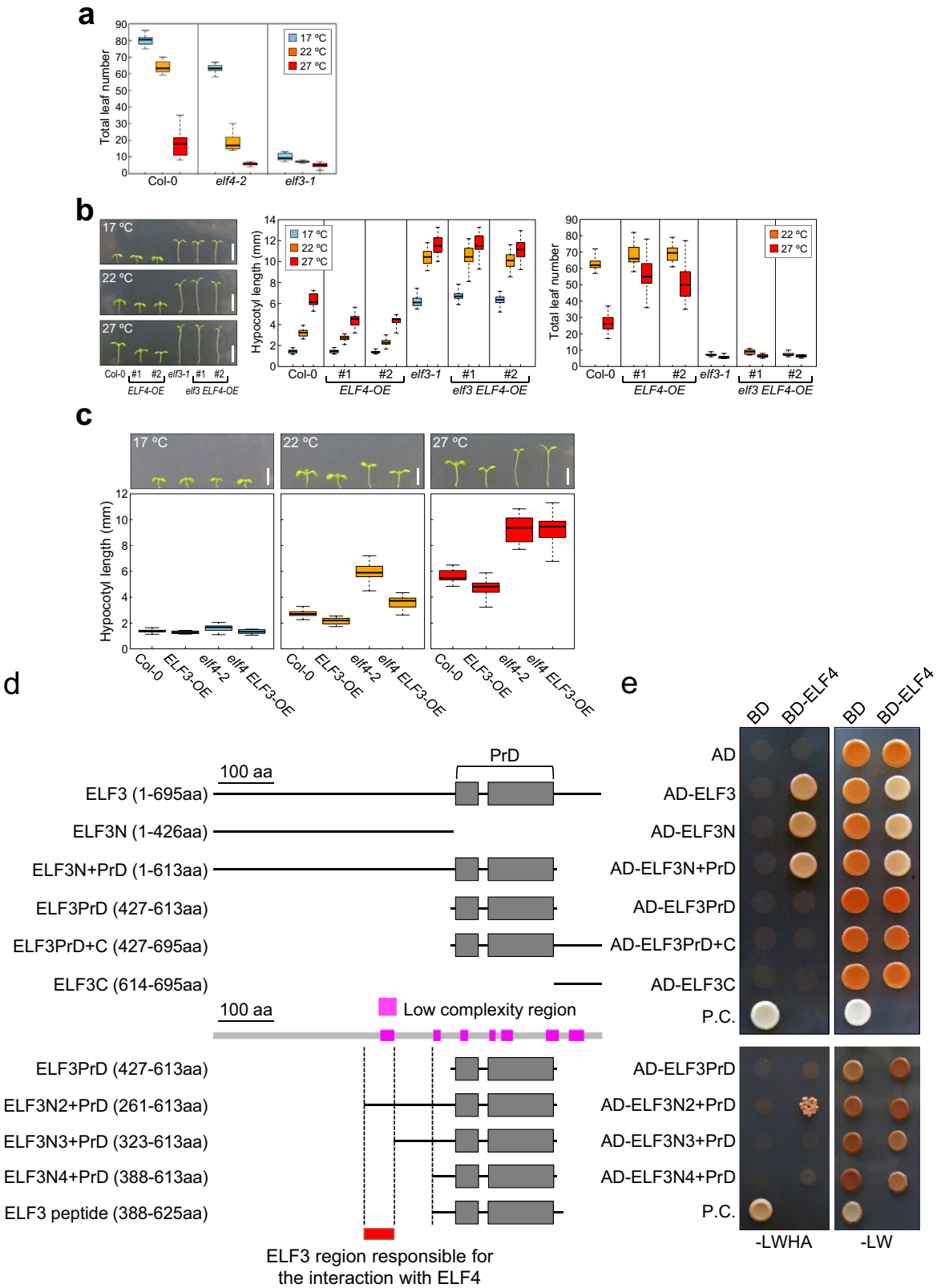
Extended Data Fig. 1 | The length of the polyQ repeat within the ELF3 PrD influences temperature responsiveness. a, Hypocotyl lengths for transgenic plants with altered ELF3 polyQ tracts grown at different temperatures. At 17 °C, *ELF3* is required to prevent hypocotyl elongation, but different polyQ lengths do not perturb ELF3 function. At 27 °C, the responsiveness of ELF3 to temperature increases with polyQ length. Q0, Q7 and Q17 refer to the length of the polyQ tract; #15 (for example) denotes a particular transgenic line. Each box is bounded by the lower and upper quartiles; the central bar represents the median; the whiskers indicate minimum and maximum values. **b**, Alignment of ELF3 amino-acid sequences from three different plant species. The region indicated by an arrow was used to create a chimeric version of *Arabidopsis* ELF3, with the ELF3 PrD replaced by the corresponding sequence of BdELF3 or StELF3. Conserved amino acids are in white in red-filled rectangles. Similar residues are in red surrounded by blue lines. Dots denote ten-amino-acid spacings.



Extended Data Fig. 2 | *ELF3* expression in the transgenic lines used here. All transgenic plants were generated by expressing *ELF3* under the control of its native promoter in *elf3-1* mutant backgrounds. The *elf3-1* phenotypes were perfectly rescued in all *ELF3* transgenic lines used here. *ELF3* transcript levels were determined by RT–qPCR. Gene-expression values were normalized to *EIF4A* expression. Data shown as means ± s.d. ($n = 3$). **a**, *ELF3pro::ELF3 elf3-1* transgenic plants without any tag sequences, used for hypocotyl-elongation and RNA-seq experiments. **b**, *ELF3pro::ELF3-FLAG elf3-1* lines, used for flowering-time measurements and ChIP–seq experiments. **c**, *ELF3pro::ELF3-GFP elf3-1* lines, used to observe ELF3-induced speckle formation in planta.

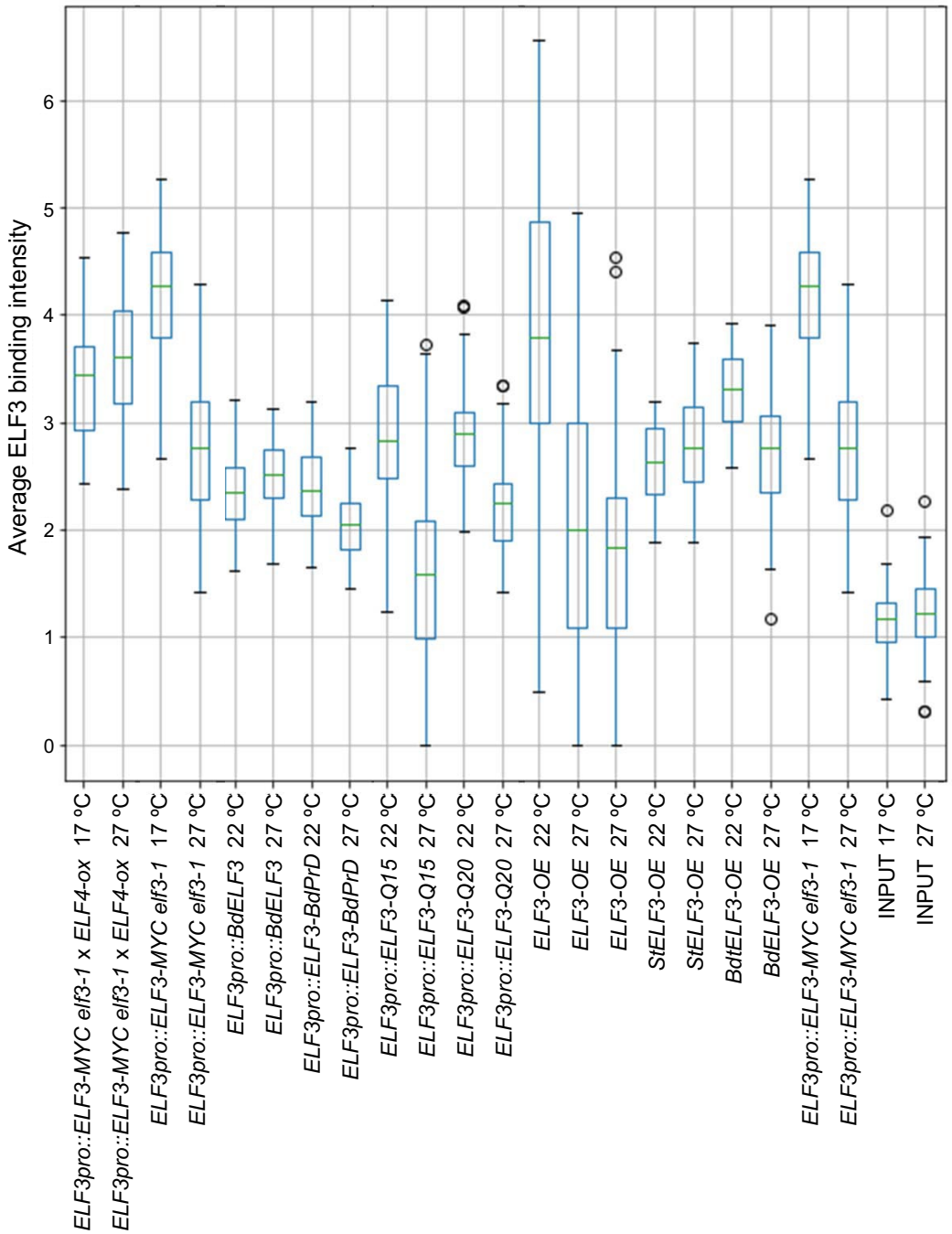


Extended Data Fig. 3 | The effects of the *elf3-1* allele in *A. thaliana* are rescued by StELF3 or BdELF3, which lack a detectable PrD. a, **b**, Transgenic *A. thaliana* plants in the *elf3-1* background, expressing different forms of *ELF3* either constitutively (from the 35S promoter; 'OE') or under the control of the endogenous *AtELF3* promoter ($ELF3_{pro}$), were grown in short-photoperiod conditions at 22 °C until bolting. **c**, Relative expression of *FT* at ZT8 was analysed by RT–qPCR. Twelve-day-old seedlings grown at different temperatures under short-photoperiod conditions (SDs) were used to analyse transcript accumulation. Data shown as means ± s.d. ($n = 3$).

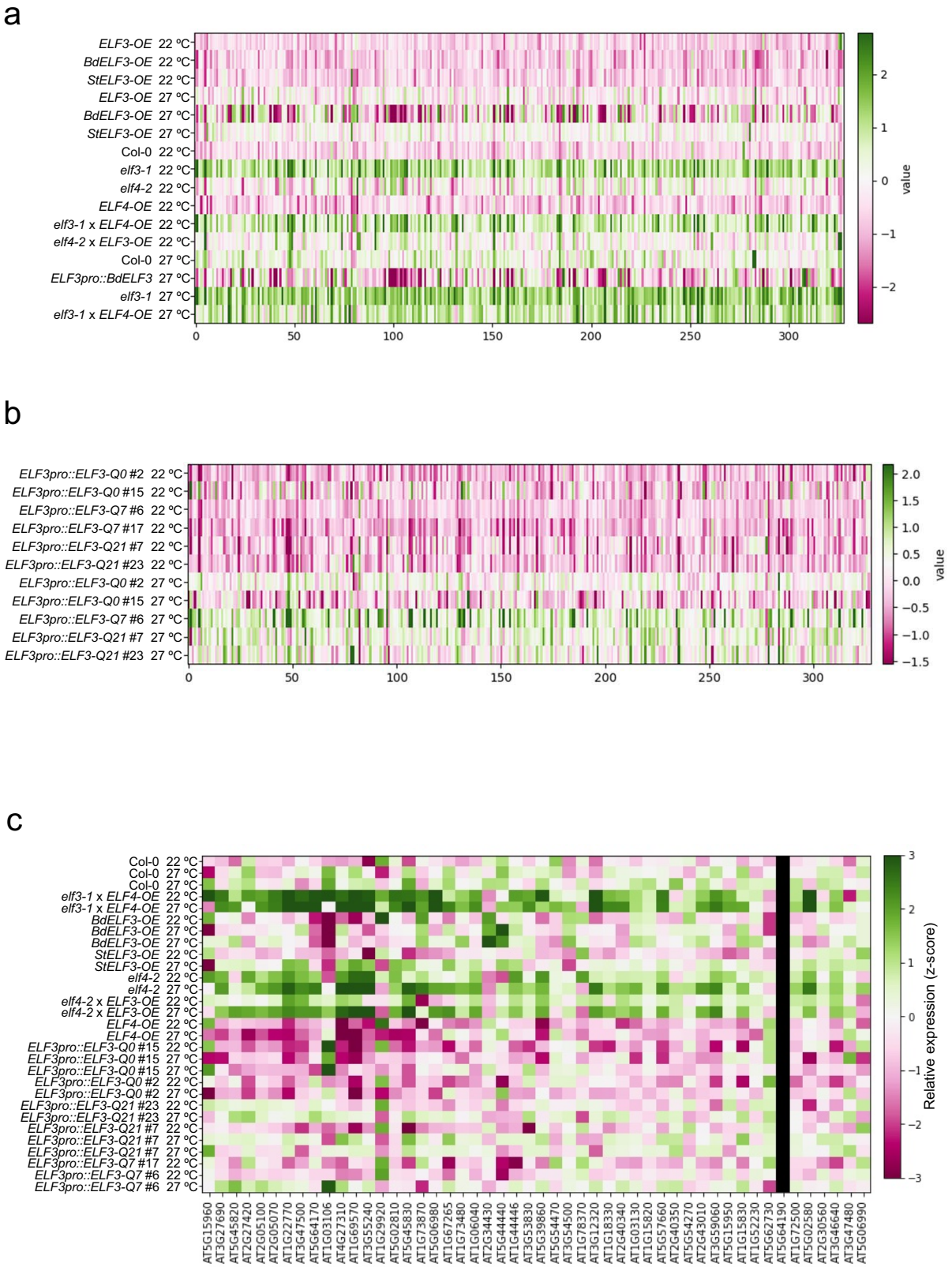


Extended Data Fig. 4 | See next page for caption.

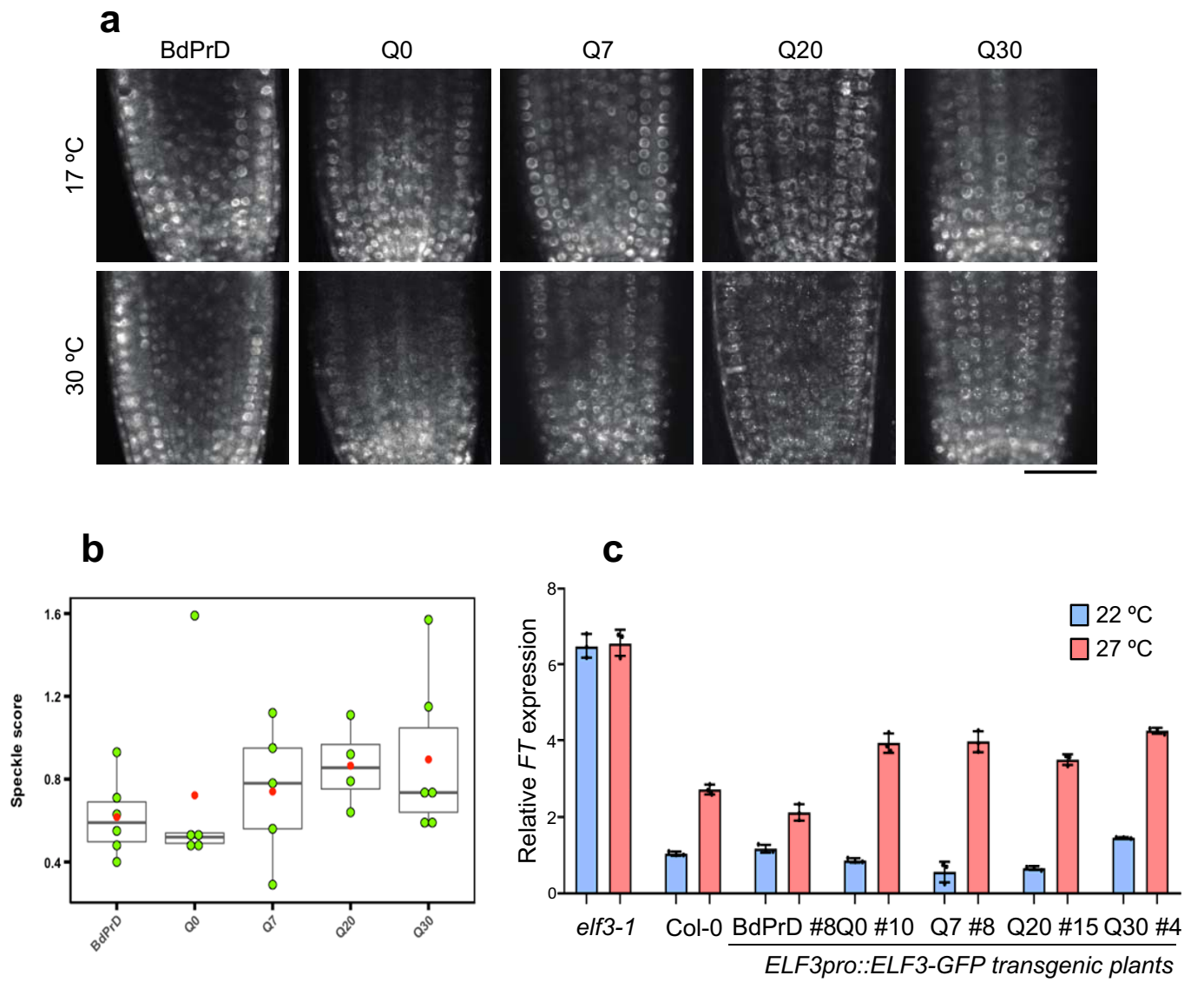
Extended Data Fig. 4 | *ELF4* is required to stabilize the activity of the evening complex at warmer temperatures. a, At lower temperatures, *ELF4* becomes dispensable for controlling flowering, but with increasing temperature, it assumes a greater role. **b**, *ELF4* overexpression greatly reduces the thermal responsiveness of both hypocotyl elongation and flowering time, and this response depends entirely on *ELF3*. **c**, At low temperatures, *ELF4* is dispensable, and *elf4-2* mutants have similar hypocotyl phenotypes to wild-type plants. As temperature increases, the role of *ELF4* becomes increasingly important, as measured by hypocotyl length. Overexpressing *ELF3* is not sufficient to change thermal responsiveness and *ELF3* overexpression has no effect in the *elf4-2* background at 27 °C, indicating that *ELF4* plays an important part at higher temperatures. In box plots, each box is bounded by the lower and upper quartiles; the central bar represents the median; the whiskers indicate minimum and maximum values. **b**, **c**, Scale bars, 5 mm. **d**, ELF3 constructs used. Numbers indicate residue positions. The domain structure of the ELF3 protein was determined using SMART protein domain annotation (http://smart.embl.de). ELF3 does not contain any specific domains except for low-complexity regions, which are regions in protein sequences that differ from the composition and complexity of most proteins with normal globular structure. **e**, Interactions of ELF3 with ELF4 in yeast cells. Cell growth on selective medium was examined. The ELF3 fragment containing a low-complexity region, which does not overlap with the PrD, is responsible for the interaction with ELF4. A soluble form of ELF3 peptide, which is used for in vitro experiments, does not include the region required for the interaction with ELF4.



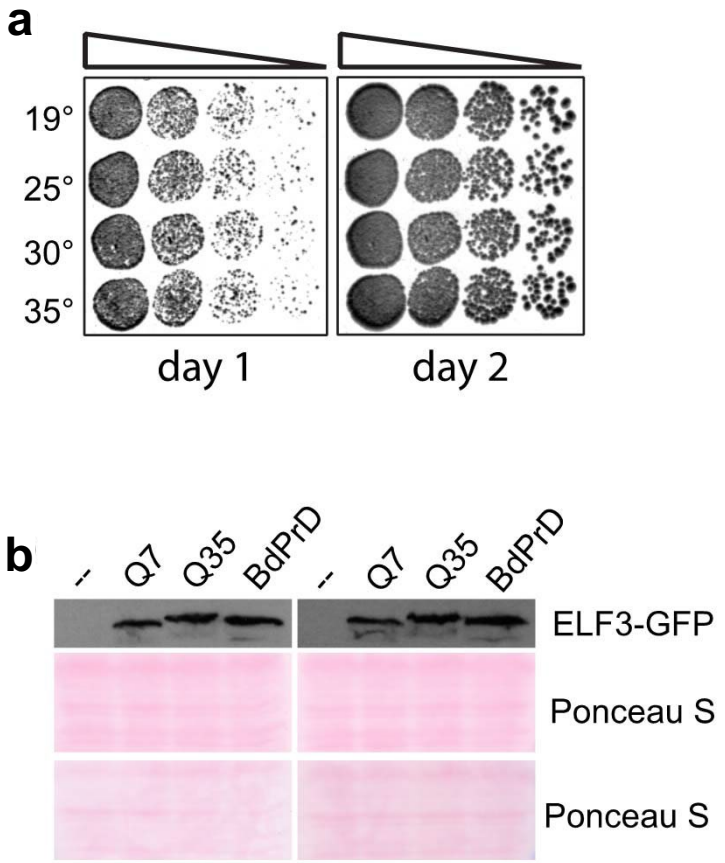
Extended Data Fig. 5 | The binding of ELF3 to target genes depends on temperature, and stabilized forms of ELF3 are less temperature responsive than wild-type ELF3. Average ELF3 ChIP–seq peak signals are measured as fold-enrichment over input (as calculated by MACS2) across multiple transgenic lines expressing the indicated ELF3 variants.



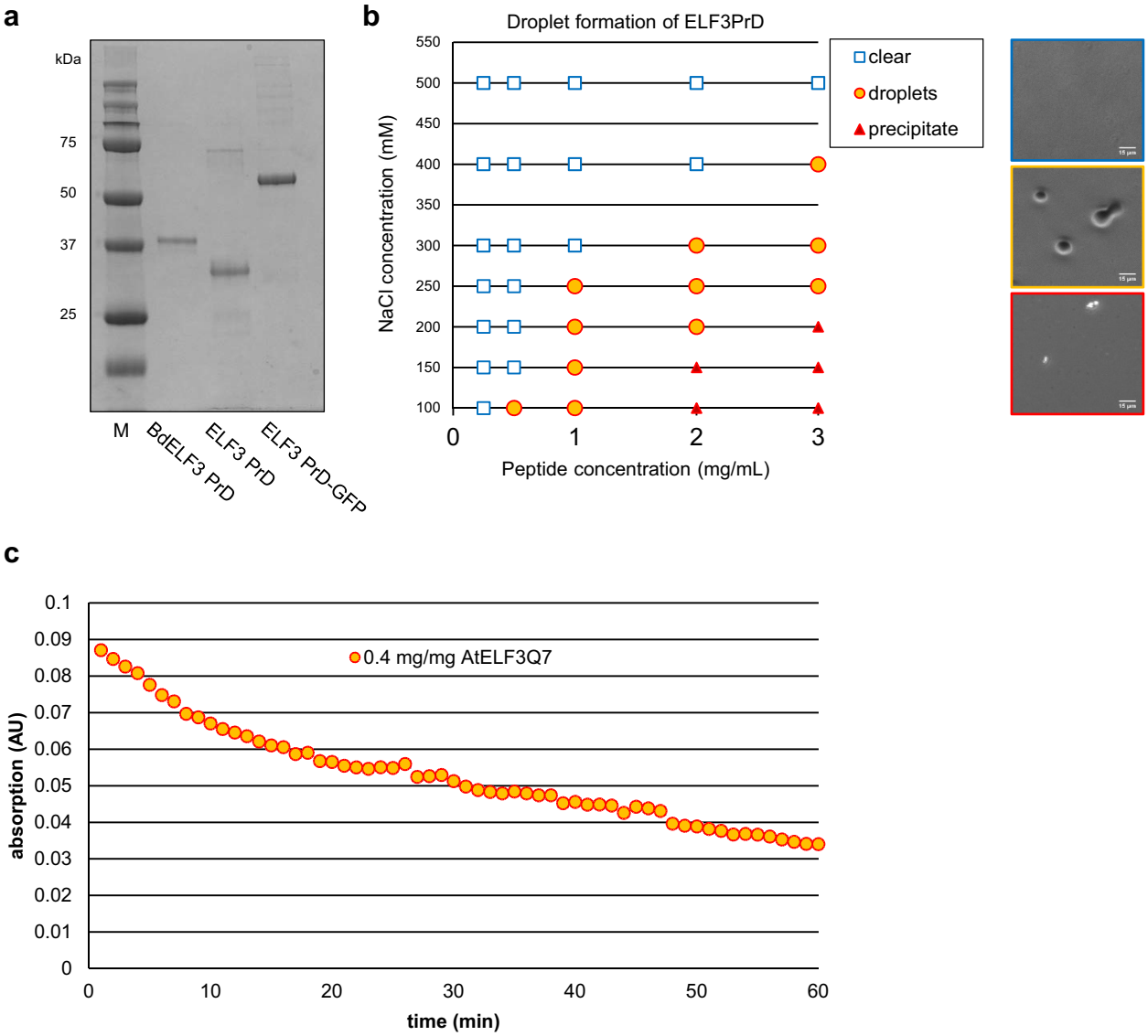
Extended Data Fig. 6 | The expression of *ELF3*-dependent genes is influenced by temperature and the PrD of ELF3. a, Effects of temperature. We analysed 325 transcripts that show *ELF3*-dependent expression in RNA-seq datasets from different genotypes at 22 °C and 27 °C. As expected, *ELF3*-dependent gene expression is generally suppressed at 22 °C (red), except in the *elf3-1* background, where genes are upregulated (green). Lines overexpressing *BdELF3* show less activation at 27 °C, consistent with their later-flowering phenotypes. Replacing just the *Arabidopsis* PrD with the corresponding region from BdELF3 (in *ELF3pro::BdELF3* at 27 °C) is sufficient to greatly reduce the upregulation of *ELF3*-dependent genes at this temperature. Upregulation of *ELF3*-dependent genes also occurs in an *elf3-1* mutant when *ELF4* is overexpressed, consistent with *ELF3* being necessary for *ELF4* action. **b**, Effects of the PrD. We analysed 325 transcripts that show *ELF3*-dependent expression in RNA-seq datasets from different polyQ genotypes at 22 °C and 27 °C. Plants expressing ELF3 with a truncated polyQ repeat (ELF3-Q0) show a reduced expression of *ELF3*-dependent genes at 27 °C, consistent with their shorter-hypocotyl phenotype. **c**, Heat map showing that ELF3-bound targets that are usually induced by shifting to 27 °C (green) become less temperature responsive in backgrounds in which ELF3 is more stable.



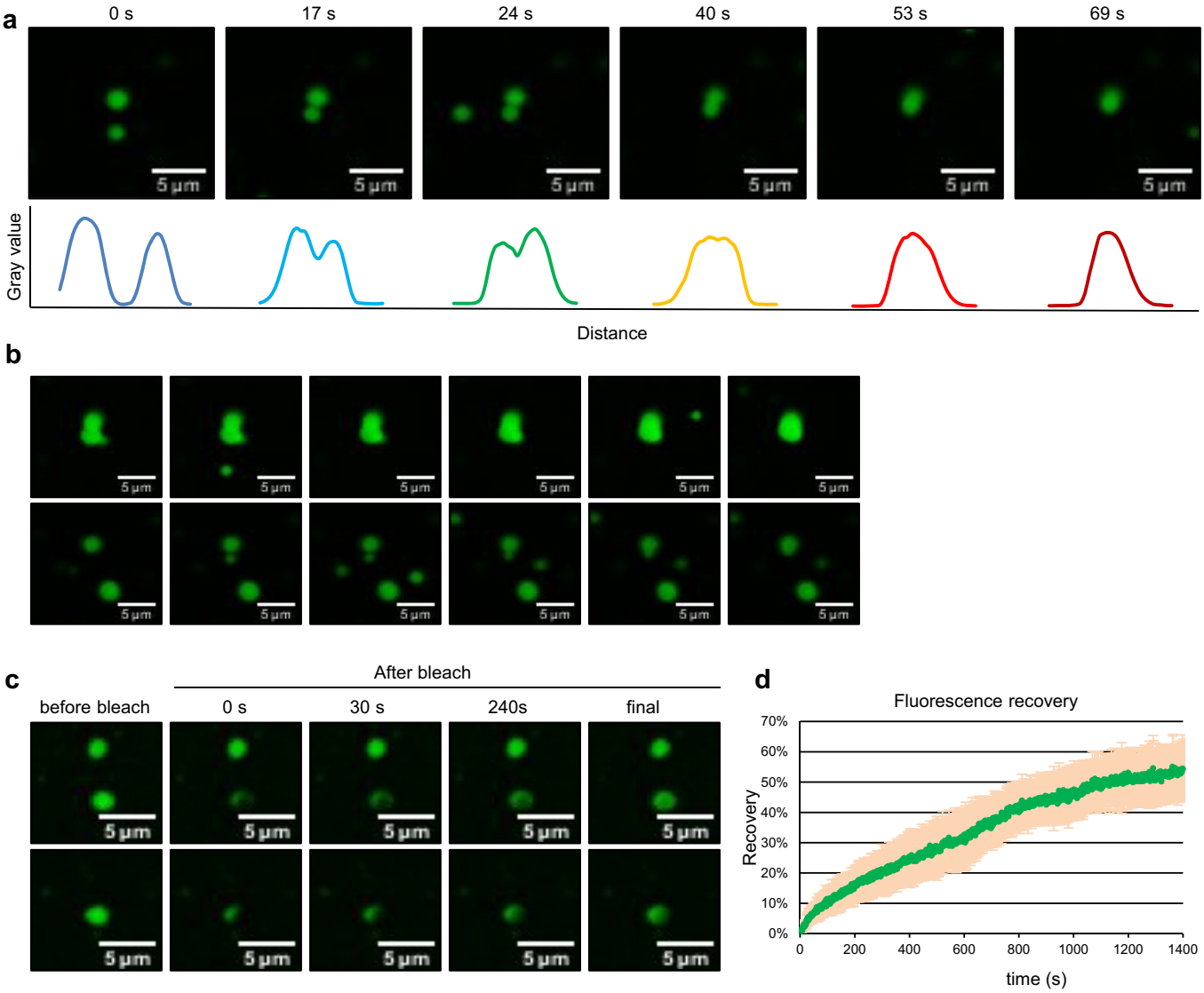
Extended Data Fig. 7 | The length of the polyQ repeat within the ELF3 PrD influences temperature-dependent speckle formation in vivo. **a**, *Arabidopsis* seedlings expressed GFP-tagged ELF3 variants with no polyQ repeat (Q0), the wild-type polyQ (Q7), a polyQ with 20 or 30 glutamines (Q20 and Q30, respectively), or the PrD replaced by the corresponding region from *B. distachyon* ELF3 (BdPrD). Seedlings were grown in short photoperiods for 7 days at 17 °C. Roots were imaged by confocal microscopy before and after incubation at 30 °C for 15 min. Scale bar, 40 µm. **b**, Quantification of the degree of speckle formation in **a**. Regions of the roots that correspond to the size of individual cells were selected, and the mean, standard deviation and maximum grey values were measured in ImageJ. We assumed that speckle formation would lead to higher grey values and that a higher frequency of speckles within the analysis region would increase the standard deviation. The lower boundary of each box indicates the 25th percentile; the median is marked by a black line within the box; and the top boundary indicates the 75th percentile. Whiskers above and below each box indicate the largest/smallest value up to 1.5 × IQR (interquartile range) from the hinge, and the red dot indicates the mean. Green dots indicate the value for each root measured. **a**, **b**, BdPrD, $n = 6$; Q0, $n = 5$; Q7, $n = 5$; Q20, $n = 4$; Q30, $n = 6$; all from two independent experiments. **c**, Relative *FT* expression in *ELF3pro::ELF3-GFP* transgenic plants. Twelve-day-old seedlings grown at different temperatures under short-photoperiod conditions were used to analyse the accumulation of *FT* transcripts at ZT8 by RT–qPCR. Results shown as means ± s.d. ($n = 3$). The effect of warm temperatures on the induction of *FT* was weak in transgenic plants containing ELF3 variants with the BdPrD.



Extended Data Fig. 8 | Yeast strains show no growth defect after incubation at temperatures used in the speckle-formation experiments, and express detectable levels of ELF3–GFP. a, Temperature shifts do not affect yeast viability. Yeast cells were grown overnight at 19 °C and shifted to the indicated temperatures for 30 min (as in the temperature shifts used for speckle inductions; Fig. 2d, e). Serial dilutions were spotted onto YPD plates and incubated at 30 °C for one or two days. **b**, Yeast cells expressing the indicated ELF3–GFP constructs (Q7, Q35 or BdPrD) or an empty vector were grown overnight in selective medium to exponential phase at 30 °C. Cells (at an optical density $(OD)_{600}$ of approximately 9) were pelleted, washed with sterile water, and lysed in 100 µl SDS-sample buffer with 0.5-mm-diameter glass beads (BioSpec Products) by two rounds of boiling for 2 min and vortexing for 30 s. Protein extracts were centrifuged at 13,000 r.p.m. for 15 min, and supernatants were analysed by western blot using anti-GFP antibody at 1:1,500 dilution (a gift from A. Peden). Western blot signals were developed using enhanced chemiluminescence (GE Healthcare).



Extended Data Fig. 9 | ELF3 PrD peptides show phase-change characteristics in vitro. a, SDS gel analysis (12% polyacrylamide) of the BdELF3 PrD, ELF3 PrD and ELF3 PrD–GFP. M, molecular-weight marker. Proteins were expressed and purified at least ten times with highly reproducible results. **b**, Phase diagram for the ELF3 PrD peptide with respect to salt and protein concentration. Examples of each phase are shown on the right. **c**, Droplet formation is dynamic, with droplets re-entering the soluble phase over time, as measured in two biological samples (mean shown) by changes in A_{340} after droplet formation is induced by dilution from a high-salt to low-salt buffer (50 mM CAPS, pH 9.7, 1 mM TCEP, 500 mM to 150 mM NaCl).



Extended Data Fig. 10 | ELF3 PrD droplets fuse. a, Fusion of two droplets over time, with intensity profiles of each droplet shown below the images. **b**, Fusion of ELF3 PrD droplets. Two examples are shown (one in each row). **c**, Example of photobleaching and recovery over time. Images were taken (left to right) before, after and at time points 30 s and 240 s post-photobleaching. **d**, FRAP recovery curves for **c**, showing means (green) ± s.d. (tan). Droplet fusions and FRAP experiments were performed five times with reproducible results.

nature research

Corresponding author(s): Philip A. Wigge

Last updated by author(s): Jul 6, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ImageJ software 1.52 ZEN software 1.8 MetaMorph software 7.8
Data analysis	https://github.com/shouldsee/polyq-figures Systat (SigmaPlot 11) Illumina sequencing reads were analysed using published open source tools as described in the materials and methods. For RNA-seq samples, Bluebee quantseq FWD analysis pipeline was used to quantify gene expression (see http://www.bluebee.com/wp-content/uploads/2017/08/QuantSeq-Data-Analysis-Pipeline-User_Guide.pdf). ChIP-seq samples were aligned with Bowtie2 after reads trimming and peaks were called with MACS2. (https://github.com/shouldsee/pipeline-chipseq-bowtie2). Code for reproducing figures are available at https://github.com/shouldsee/polyq-figures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability Statement:
Sequencing data for gene expression analysis (RNA-seq) and protein-DNA interactions (ChIP-seq) have been deposited in the publicly available databasej GEO, Accession Code GSE137264 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137264). The raw data used in this study are available at: https://osf.io/fn5um/

Code Availability:
The code to produce the figures from the processed files is available at https://github.com/shouldsee/polyq-figures . To enable easier browsing, a static site is hosted at https://shouldsee.github.io/polyq-figures . The inhouse pipeline for mapping is available at https://github.com/shouldsee/synoBio .

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen that are widely used in the field, for example see Jung et al., Science 354:886. No statistical tests were used to predetermine sample size.
Data exclusions	No results or data were excluded from this study.
Replication	All replicates were successful, and nothing has been excluded from this study. Values of n (counting number) and numbers of biological replicates are indicated in the figure legends.
Randomization	Plants were grown in carefully randomised growth room conditions, and pots rotated between trays to ensure all plants received identical conditions. Seedlings were grown on sterile MS-agar medium on tissue culture plates. They were randomly mixed and processed.
Blinding	Investigators were blinded during both automated data analysis and manual counting.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Anti-c-Myc agarose affinity gel antibody (Sigma, A7470), Anti-HA-Agarose (Sigma, A2095) or Anti-Flag M2 Affinity Gel (Sigma, A2220)

were used for ChIP-seq experiments. GFP was detected with polyclonal anti-GFP, a gift from A. Peden (Univ. Sheffield). All ChIP experiments were performed with resin/agarose volumes of 100 µl of resuspended resin/agarose antibody (anti-c-Myc, anti-HA and anti-Flag) (equivalent to 50 µl of settled resin/agarose). The anti-GFP was used at 1:1500 dilution for western blotting.

Validation

Anti-c-Myc agarose affinity gel antibody (Sigma, A7470), https://simgaalrdrich.com/catalog/product/sigma/a7470
Anti-HA-Agarose (Sigma, A2095), https://simgaalrdrich.com/catalog/product/sigma/a2095
Anti-Flag M2 Affinity Gel (Sigma, A2220), https://simgaalrdrich.com/catalog/product/sigma/a2220
Anti-GFP: Gordon et al PloS Genetics 2017: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006698

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Yeast cells (RS453 MATα ade2-1 his3-11,15 ura3-52 leu2-3112 trp1-1, URA3::YIplac211-SEC63-mCherry)
doi: 10.1091/mbc.E15-03-0173

Authentication

Western blotting.

Mycoplasma contamination

N/A

Commonly misidentified lines
(See ICLAC register)

N/A

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

RNASeq and CHIPSeq data on GSE137264: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137264

Files in database submission

Raw fastq files and aligned bam files: deposited in the related SRA entry (SRA files are not viewable before making public)
*_RPKM.bw: RPKM-normalised bigwig track at 10bp resolution
*_.narrowPeak: containing MACS2-called peaks. as described in methods.
(Note that GEO added a prefix to the submitted files and is not included here)

176CS11.supp.1491-22C-ZT10_S11_Ath-TAIR10_peaks.narrowPeak
176CS11.supp.1491-22C-ZT10_S11_Ath-TAIR10_RPKM.bw
176CS12.supp.1491-27C-ZT10_S12_Ath-TAIR10_peaks.narrowPeak
176CS12.supp.1491-27C-ZT10_S12_Ath-TAIR10_RPKM.bw
176CS17.supp.1490-22C-ZT10_S17_Ath-TAIR10_peaks.narrowPeak
176CS17.supp.1490-22C-ZT10_S17_Ath-TAIR10_RPKM.bw
176CS18.supp.1490-27C-ZT10_S18_Ath-TAIR10_peaks.narrowPeak
176CS18.supp.1490-27C-ZT10_S18_Ath-TAIR10_RPKM.bw
176CS19.supp.35ELF3-HA-22C-ZT10_S19_Ath-TAIR10_peaks.narrowPeak
176CS19.supp.35ELF3-HA-22C-ZT10_S19_Ath-TAIR10_RPKM.bw
176CS1.supp.1487-22C-ZT10_S1_Ath-TAIR10_peaks.narrowPeak
176CS1.supp.1487-22C-ZT10_S1_Ath-TAIR10_RPKM.bw
176CS20.supp.35ELF3-HA-27C-ZT10_S20_Ath-TAIR10_peaks.narrowPeak
176CS20.supp.35ELF3-HA-27C-ZT10_S20_Ath-TAIR10_RPKM.bw
176CS21.supp.INPUT-379_S21_Ath-TAIR10_peaks.narrowPeak
176CS21.supp.INPUT-379_S21_Ath-TAIR10_RPKM.bw
176CS22.supp.INPUT-380_S22_Ath-TAIR10_peaks.narrowPeak
176CS22.supp.INPUT-380_S22_Ath-TAIR10_RPKM.bw
176CS3.supp.1488-22C-ZT10_S3_Ath-TAIR10_peaks.narrowPeak
176CS3.supp.1488-22C-ZT10_S3_Ath-TAIR10_RPKM.bw
176CS4.supp.1488-27C-ZT10_S4_Ath-TAIR10_peaks.narrowPeak
176CS4.supp.1488-27C-ZT10_S4_Ath-TAIR10_RPKM.bw
176CS5.supp.4195-22C-ZT10_S5_Ath-TAIR10_peaks.narrowPeak
176CS5.supp.4195-22C-ZT10_S5_Ath-TAIR10_RPKM.bw
176CS6.supp.4195-27C-Zt10_S6_Ath-TAIR10_peaks.narrowPeak
176CS6.supp.4195-27C-Zt10_S6_Ath-TAIR10_RPKM.bw
176CS7.supp.4196-22C-ZT10_S7_Ath-TAIR10_peaks.narrowPeak
176CS7.supp.4196-22C-ZT10_S7_Ath-TAIR10_RPKM.bw
176CS8.supp.4196-27C-ZT10_S8_Ath-TAIR10_peaks.narrowPeak
176CS8.supp.4196-27C-ZT10_S8_Ath-TAIR10_RPKM.bw

182CS24.supp.35SELF3-27C-HA_S24_Ath-TAIR10_peaks.narrowPeak
182CS24.supp.35SELF3-27C-HA_S24_Ath-TAIR10_RPKM.bw
189CS10.supp.gELF3myc-17C_S10_peaks.narrowPeak
189CS10.supp.gELF3myc-17C_S10_RPKM.bw
189CS11.supp.gELF3myc-27C_S11_RPKM.bw
189CS16.supp.1469-gELF3-myc-elf3-1xELF4-ox-ZT10-17C_S16_peaks.narrowPeak
189CS16.supp.1469-gELF3-myc-elf3-1xELF4-ox-ZT10-17C_S16_RPKM.bw
189CS17.supp.1469-gELF3-myc-elf3-1xELF4-ox-ZT10-27C_S17_peaks.narrowPeak
189CS17.supp.1469-gELF3-myc-elf3-1xELF4-ox-ZT10-27C_S17_RPKM.bw
192CS17.supp.ELF3myc-17C-ZT10_S17_Ath-TAIR10_peaks.narrowPeak
192CS17.supp.ELF3myc-17C-ZT10_S17_Ath-TAIR10_RPKM.bw
192CS18.supp.ELF3myc-27C-ZT10_S18_Ath-TAIR10_peaks.narrowPeak
192CS18.supp.ELF3myc-27C-ZT10_S18_Ath-TAIR10_RPKM.bw
192CS1.supp.1487-17-ZT10_S1_Ath-TAIR10_peaks.narrowPeak
192CS1.supp.1487-17-ZT10_S1_Ath-TAIR10_RPKM.bw
192CS2.supp.1487-27-ZT10_S2_Ath-TAIR10_peaks.narrowPeak
192CS2.supp.1487-27-ZT10_S2_Ath-TAIR10_RPKM.bw
192CS3.supp.1488-17-ZT10_S3_Ath-TAIR10_peaks.narrowPeak
192CS3.supp.1488-17-ZT10_S3_Ath-TAIR10_RPKM.bw
192CS4.supp.1488-27-ZT10_S4_Ath-TAIR10_peaks.narrowPeak
192CS4.supp.1488-27-ZT10_S4_Ath-TAIR10_RPKM.bw

Genome browser session
(e.g. UCSC)

NA

Methodology

Replicates

To ensure the observation of temperature-sensitive peaks is reproducible, two independent experiments (189CS10+189CS11 and 192CS17+192CS18) were cross-compared.

Sequencing depth

sample_accession=186CS12
number_of_reads=10.40M
number_of_uniq_mapped_reads=6.31M
rawfile_readlengths=75
rawfile_is_paired=single

sample_accession=192CS9
number_of_reads=32.75M
number_of_uniq_mapped_reads=15.39M
rawfile_readlengths=75
rawfile_is_paired=single

sample_accession=192CS10
number_of_reads=21.28M
number_of_uniq_mapped_reads=6.28M
rawfile_readlengths=75
rawfile_is_paired=single

sample_accession=192CS11
number_of_reads=18.51M
number_of_uniq_mapped_reads=8.10M
rawfile_readlengths=75
rawfile_is_paired=single

sample_accession=192CS12
number_of_reads=20.31M
number_of_uniq_mapped_reads=8.36M
rawfile_readlengths=75
rawfile_is_paired=single

Antibodies

Anti-c-Myc agarose affinity gel antibody (Sigma, A7470), https://simgaalrdrich.com/catalog/product/sigma/a7470
Anti-HA-Agarose (Sigma, A2095), https://simgaalrdrich.com/catalog/product/sigma/a2095
Anti-Flag M2 Affinity Gel (Sigma, A2220), https://simgaalrdrich.com/catalog/product/sigma/a2220

Peak calling parameters

For each treated ChIP-Seq library, peaks were called against a control 176CS21 (INPUT genomic DNA) using MACS2 with argument "--keep-dup 1 -p 0.1".

Data quality

sample_accession=176CS1
number_of_peaks_below_5%FDR=6361
number_of_peaks_above_5fold_enrichment=48

sample_accession=176CS3
number_of_peaks_below_5%FDR=1487
number_of_peaks_above_5fold_enrichment=20

sample_accession=176CS4
number_of_peaks_below_5%FDR=887
number_of_peaks_above_5fold_enrichment=7

sample_accession=176CS5
number_of_peaks_below_5%FDR=468
number_of_peaks_above_5fold_enrichment=17

sample_accession=176CS6
number_of_peaks_below_5%FDR=9953
number_of_peaks_above_5fold_enrichment=85

sample_accession=176CS7
number_of_peaks_below_5%FDR=7452
number_of_peaks_above_5fold_enrichment=109

sample_accession=176CS8
number_of_peaks_below_5%FDR=1612
number_of_peaks_above_5fold_enrichment=109

sample_accession=176CS11
number_of_peaks_below_5%FDR=1965
number_of_peaks_above_5fold_enrichment=77

sample_accession=176CS12
number_of_peaks_below_5%FDR=1239
number_of_peaks_above_5fold_enrichment=10

sample_accession=176CS17
number_of_peaks_below_5%FDR=8143
number_of_peaks_above_5fold_enrichment=449

sample_accession=176CS18
number_of_peaks_below_5%FDR=3829
number_of_peaks_above_5fold_enrichment=1453

sample_accession=176CS19
number_of_peaks_below_5%FDR=5377
number_of_peaks_above_5fold_enrichment=2064

sample_accession=176CS20
number_of_peaks_below_5%FDR=1809
number_of_peaks_above_5fold_enrichment=682

sample_accession=176CS21
number_of_peaks_below_5%FDR=0
number_of_peaks_above_5fold_enrichment=0

sample_accession=176CS22
number_of_peaks_below_5%FDR=30
number_of_peaks_above_5fold_enrichment=0

sample_accession=182CS24
number_of_peaks_below_5%FDR=3297
number_of_peaks_above_5fold_enrichment=1117

sample_accession=189CS10
number_of_peaks_below_5%FDR=4519
number_of_peaks_above_5fold_enrichment=130

sample_accession=189CS11
number_of_peaks_below_5%FDR=NA
number_of_peaks_above_5fold_enrichment=NA

sample_accession=189CS16
number_of_peaks_below_5%FDR=1722
number_of_peaks_above_5fold_enrichment=50

sample_accession=189CS17
number_of_peaks_below_5%FDR=1684
number_of_peaks_above_5fold_enrichment=69

sample_accession=192CS1
number_of_peaks_below_5%FDR=6274
number_of_peaks_above_5fold_enrichment=11

sample_accession=192CS17
number_of_peaks_below_5%FDR=2832
number_of_peaks_above_5fold_enrichment=114

sample_accession=192CS18
number_of_peaks_below_5%FDR=586
number_of_peaks_above_5fold_enrichment=29

sample_accession=192CS2
number_of_peaks_below_5%FDR=1590
number_of_peaks_above_5fold_enrichment=10

sample_accession=192CS3
number_of_peaks_below_5%FDR=444
number_of_peaks_above_5fold_enrichment=7

sample_accession=192CS4
number_of_peaks_below_5%FDR=1262
number_of_peaks_above_5fold_enrichment=11

Software

After MACS2 was used to call crude peaks, inhouse Python2 code was applied to extract target peaks and producing pile-ups. See github for detail.

Article

Distinct viral reservoirs in individuals with spontaneous control of HIV-1

https://doi.org/10.1038/s41586-020-2651-8

Received: 2 October 2019

Accepted: 15 July 2020

Published online: 26 August 2020

Check for updates

Chenyang Jiang[1,2,15], Xiaodong Lian[1,2,15], Ce Gao[1,15], Xiaoming Sun[1], Kevin B. Einkauf[1,2], Joshua M. Chevalier[1,2], Samantha M. Y. Chen[1], Stephane Hua[1], Ben Rhee[1,2], Kaylee Chang[1], Jane E. Blackmer[1], Matthew Osborn[1], Michael J. Peluso[3], Rebecca Hoh[3], Ma Somsouk[3], Jeffrey Milush[3], Lynn N. Bertagnolli[4], Sarah E. Sweet[4], Joseph A. Varriale[4], Peter D. Burbelo[5], Tae-Wook Chun[6], Gregory M. Laird[7], Erik Serrao[8,9], Alan N. Engelman[8,9], Mary Carrington[1,10], Robert F. Siliciano[4,11], Janet M. Siliciano[4,11], Steven G. Deeks[3], Bruce D. Walker[1,11,12,13], Mathias Lichterfeld[1,2,14] & Xu G. Yu[1,2] ✉

Sustained, drug-free control of HIV-1 replication is naturally achieved in less than 0.5% of infected individuals (here termed 'elite controllers'), despite the presence of a replication-competent viral reservoir[1]. Inducing such an ability to spontaneously maintain undetectable plasma viraemia is a major objective of HIV-1 cure research, but the characteristics of proviral reservoirs in elite controllers remain to be determined. Here, using next-generation sequencing of near-full-length single HIV-1 genomes and corresponding chromosomal integration sites, we show that the proviral reservoirs of elite controllers frequently consist of oligoclonal to near-monoclonal clusters of intact proviral sequences. In contrast to individuals treated with long-term antiretroviral therapy, intact proviral sequences from elite controllers were integrated at highly distinct sites in the human genome and were preferentially located in centromeric satellite DNA or in Krüppel-associated box domain-containing zinc finger genes on chromosome 19, both of which are associated with heterochromatin features. Moreover, the integration sites of intact proviral sequences from elite controllers showed an increased distance to transcriptional start sites and accessible chromatin of the host genome and were enriched in repressive chromatin marks. These data suggest that a distinct configuration of the proviral reservoir represents a structural correlate of natural viral control, and that the quality, rather than the quantity, of viral reservoirs can be an important distinguishing feature for a functional cure of HIV-1 infection. Moreover, in one elite controller, we were unable to detect intact proviral sequences despite analysing more than 1.5 billion peripheral blood mononuclear cells, which raises the possibility that a sterilizing cure of HIV-1 infection, which has previously been observed only following allogeneic haematopoietic stem cell transplantation[2,3], may be feasible in rare instances.

Individuals with untreated HIV-1 infections who durably control HIV-1 replication below the threshold of detection of commercial viral load assays (here termed 'elite controllers') may represent the closest possible approximation to a natural cure of HIV-1 infection[1]. Previous studies have linked elite HIV-1 control to specific variations in the human HLA class I gene locus[4], and to the presence of highly functional cellular immune responses[5] that have stronger abilities to kill virus-infected cells[5], target mutationally constrained epitopes[6] and limit viral escape[7]. Although the persistence of small, replication-competent proviral reservoirs has been documented in elite controllers[8,9], the characteristics and possible distinguishing features of reservoir cells in this specific group of individuals remain poorly defined.

We used full-length individual provirus sequencing (FLIP-seq)[10] to profile the proviral reservoir landscape at single-genome resolution of a large cohort of elite controllers who maintained undetectable HIV-1 plasma viral loads for a median of 9 years (range, 1–24 years) based on commercially available PCR assays. A reference cohort of individuals with HIV-1 infections who were treated with suppressive antiretroviral

[1]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [2]Infectious Disease Division, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, University of California at San Francisco, San Francisco, CA, USA. [4]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Dental Clinical Research Core, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, USA. [6]National Institute of Allergies and Infectious Diseases, Bethesda, MD, USA. [7]Accelevir Diagnostics, Baltimore, MD, USA. [8]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Medicine, Harvard Medical School, Boston, MA, USA. [10]Basic Science Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [11]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [12]Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [13]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [14]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]These authors contributed equally: Chenyang Jiang, Xiaodong Lian, Ce Gao. ✉e-mail: xyu@mgh.harvard.edu

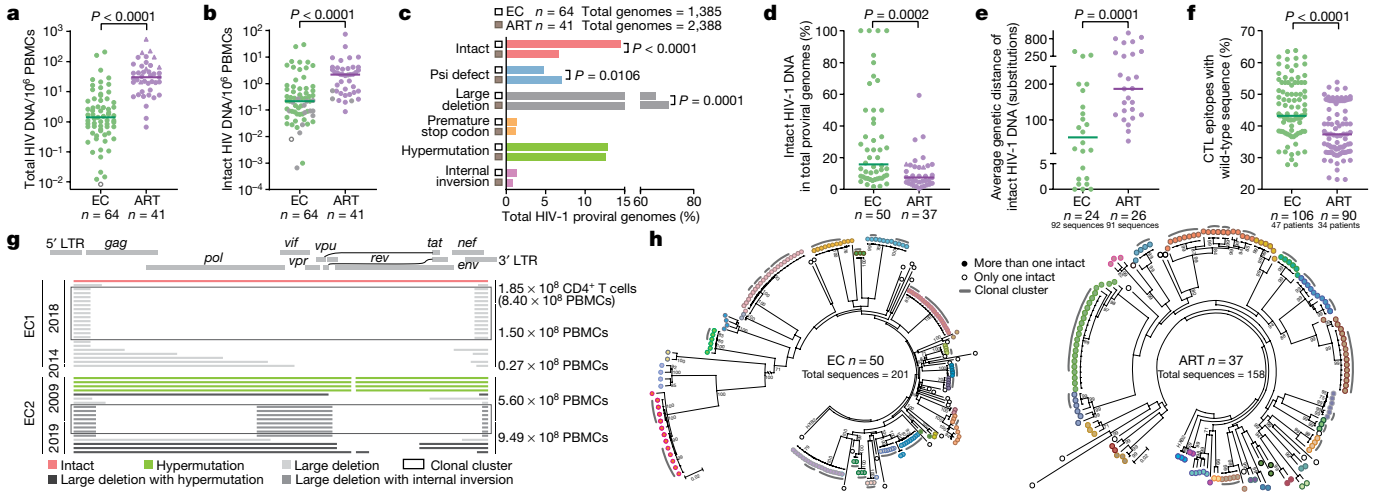


Fig. 1 | Proviral reservoir landscape in HIV-1 elite controllers. a, **b**, Relative frequencies of total (**a**) and near-full-length intact (**b**) HIV-1 DNA sequences in elite controllers (EC) and ART-treated individuals (ART). Grey symbols, limit of detection (expressed as 1 copy/total number of analysed cells without target identification). Circles, proviral sequences obtained from unfractionated PBMCs; triangles, proviral sequences retrieved from isolated CD4$^+$ T cells and normalized to the number of PBMCs. Open circles show the Berlin patient. **c**, Proportions of proviral sequences that have an intact genome or display defined structural defects among all proviral genomes. Psi, packaging signal. **d**, Proportion of genome-intact proviral sequences among all proviral genomes from each study participant. Only individuals for whom at least one genome-intact proviral sequence was detected are shown. **e**, Average genetic distance between distinct genome-intact proviral sequences obtained from each study participant. Participants with at least two detectable genome-intact proviral sequences are included. **f**, Proportion of optimal CTL epitopes (restricted by autologous HLA class I isotypes) with wild-type clade B consensus sequences within a given clade B genome-intact proviral sequence. Each dot represents data from one genome-intact proviral sequence. Clonal sequences are counted once. **g**, All proviral HIV-1 sequences isolated from EC1 and EC2. Dates of sample collection are indicated on the left; numbers of cells analysed are indicated on the right. Open boxes indicate clonal clusters. **h**, Circular maximum-likelihood phylogenetic trees for all genome-intact proviral sequences from elite controllers and ART-treated individuals. HXB2, reference HIV-1 sequence. Dots with the same colours indicate genome-intact proviral sequences that were detected in the same individuals. Clonal sequences, defined by complete sequence identity, are indicated by grey arches. Bootstrap analysis with 1,000 replicates was performed to assign confidence to tree nodes; bootstrap support values >70% are shown in the trees. Two-tailed Mann–Whitney *U*-tests were used for data shown in **a**, **b**, **d**–**f**; false-discovery rate (FDR)-adjusted two-tailed Fisher's exact tests were used for data shown in **c**.

therapy (ART) for a median of 9 years (range, 2–19 years) was recruited for comparative purposes (Extended Data Table 1). Collectively, our analysis of a large number of individual HIV-1 proviral genomes (n = 1,385 from 64 elite controllers and n = 2,388 from 41 ART-treated individuals) demonstrated that the median number of proviral amplification products (intact and defective) per person was significantly lower in elite controllers relative to ART-treated individuals (Fig. 1a). Frequencies of near-full-length proviral sequences with intact genomes that did not contain defined lethal sequence defects were also markedly reduced in elite controllers, although their quantitative spectrum varied considerably (Fig. 1b). Of note, genome-intact proviral sequences made up a significantly larger proportion of all proviral sequences in elite controllers at both the cohort level (Fig. 1c) and the per-study participant level (Fig. 1d) compared to ART-treated individuals; in four elite controllers, genome-intact proviral sequences accounted for 100% of the detected proviral species. Intra-individual diversity in the proviral sequences, determined by pair-wise comparisons of all genome-intact proviral sequences within a given study participant, was smaller in elite controllers (Fig. 1e and Extended Data Fig. 1a). Notably, within genome-intact proviral sequences from elite controllers, optimal epitope sequences of cytotoxic T lymphocytes (CTLs) restricted by autologous HLA class I isotypes displayed more limited evidence of mutational escape (Fig. 1f and Extended Data Fig. 1c–f). These data suggest that genome-intact proviral sequences from elite controllers were seeded early in the disease process and persisted long-term.

For a more in-depth analysis of the structure of the proviral reservoir, we initially focused on two elite controllers for whom no genome-intact proviral sequences were observed in our initial analysis. For EC1—an individual who had maintained drug-free HIV-1 control for a recorded time of 12 years with only one documented episode of viraemia of 56 HIV-1 RNA copies per ml out of 23 viral load tests that spanned this period (Extended Data Fig. 2)—we increased the number of analysed peripheral blood mononuclear cells (PBMCs) to the limit of available cells and found a single genome-intact proviral sequence in a total of 1.02 billion PBMCs analysed; 21 defective proviruses, many of which belonged to a sequence-identical cluster (Fig. 1g), were also detected. For EC2—who had a single documented episode of 93 HIV-1 RNA copies per ml in 39 viral load tests that spanned more than 24 years of follow-up without ART (Extended Data Fig. 2)—we did not detect a single genome-intact proviral sequence in more than 1.5 billion PBMCs, although 19 defective proviral species, including near-full-length sequences with lethal hypermutations, were observed, which clearly indicates that this individual had been infected with HIV-1 in the past (Fig. 1g). Members of a sequence-identical cluster of defective proviral sequences with large deletions were noted in samples that had been collected in 2009 and in 2019 from EC2, demonstrating the durability of a clonal cell population that contains this sequence.

Moreover, a subsequent quantitative viral outgrowth assay (qVOA) with 340 million resting CD4$^+$ T cells isolated from approximately 1 billion PBMCs (collected in 2019), and an additional qVOA that included 41 million total CD4$^+$ T cells isolated from 158.5 million PBMCs (collected in 2009) did not retrieve a single replication-competent viral species. The recently developed intact proviral DNA assay did not find evidence of genome-intact proviral sequences in 14 million resting CD4$^+$ T cells, but confirmed the presence of defective HIV-1 DNA sequences (Extended Data Fig. 1b). In addition, an analysis of 7.72 million gut cells collected by colonoscopy from the rectum (2.08 million CD45$^+$ mononuclear cells and 2.30 million CD45$^-$ cells) and terminal ileum (1.99 million CD45$^+$ mononuclear cells and 1.35 million CD45$^-$ cells) by FLIP-seq did not reveal any intact or defective proviruses in samples from EC2.

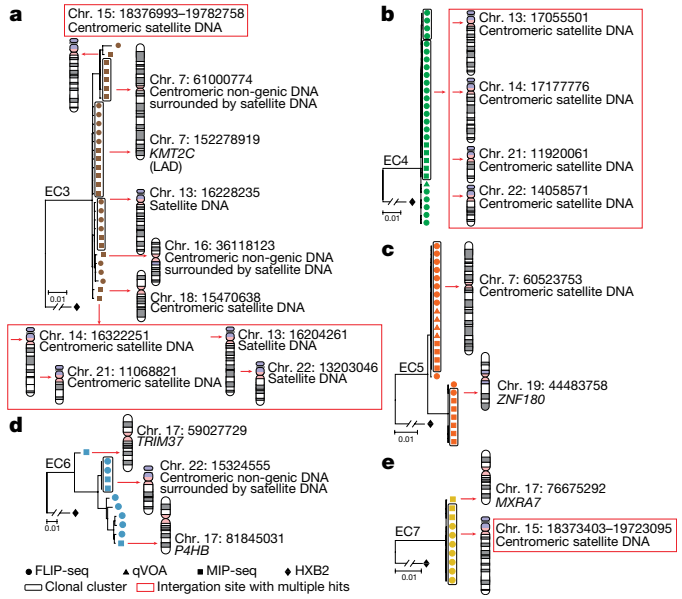


Fig. 2 | Increased frequency of genome-intact proviral sequences integrated in centromeric satellite DNA in elite controllers. a–e, Linear maximum-likelihood phylogenetic trees for genome-intact proviral sequences from five elite controllers are shown. Coordinates and relative positioning of integration sites are depicted; genes that contain integration sites are listed. Clonal genome-intact proviral sequences, defined by identical proviral sequences and identical corresponding integration sites, are highlighted in black boxes. Red boxes reflect multi-hit integration sites that cannot be definitively mapped to one particular genomic location owing to their position in repetitive centromeric satellite DNA that is present in multiple regions of the human genome. LAD, lamina-associated domain.

To our knowledge, the absence of genome-intact proviral sequences in such extremely large numbers of analysed cells has been documented only in the 'Berlin patient' who underwent an allogeneic haematopoietic stem cell transplantation from a donor who was homozygous for CCR5Δ32[2]; this resulted in what is widely considered a sterilizing cure of HIV-1 infection. Indeed, we did not retrieve any intact or defective proviral sequences using FLIP-seq in an analysis of 113 million PBMCs from the Berlin patient (collected in 2017 and 2018) (Fig. 1a, b). Although the logic of scientific discovery[11] does not allow us to confirm that EC2 has achieved a sterilizing cure of HIV-1 infection through natural immune-mediated mechanisms, it is notable that we have failed to falsify this hypothesis, despite analysing large amounts of cells with a range of complementary, highly sensitive detection techniques.

We next performed a phylogenetic analysis of all genome-intact proviral sequences obtained from 50 elite controllers and 37 ART-treated individuals. In both groups, we readily observed large clusters of sequences that were completely identical over entire analysed viral genomes (Fig. 1h), strongly suggesting that they originate from clonally expanded HIV-1-infected cells that passed on identical copies of genome-intact proviral sequences during cell divisions. The proportions of these genome-intact proviral sequences derived from clonally expanded cells were significantly higher in elite controllers than in ART-treated individuals (Extended Data Fig. 1g, h). A number of these sequences were also retrieved in qVOAs, indicating that these genome-intact proviral sequences are fully replication-competent (Figs. 2, 3).

For a detailed analysis of the viral reservoir landscape in elite controllers, we focused on eleven elite controllers (EC3–EC13), in whom large clusters of identical genome-intact proviral sequences were detected and from whom sufficient numbers of cells were available. We frequently observed oligoclonal, and sometimes almost monoclonal, compositions of the entire intact proviral reservoir landscape in cells from these individuals (Figs. 2, 3 and Extended Data Fig. 3). Notably, such a narrowly focused configuration of the viral reservoir that consists of few distinct genome-intact proviral sequences but displays relatively large expansions of identical clones of genome-intact proviral sequences is compatible with very low—if any—levels of ongoing viral replication in these elite controllers. This structure of the viral reservoir is atypical relative to the more-diverse spectrum of genome-intact proviral sequences that have previously been described for long-term ART-treated individuals[10,12]. Instead, the landscape of the viral reservoir of EC3–EC13 is more similar to the oligoclonal structure of the viral reservoir of genome-intact proviral sequences that are typically observed in individuals with chronic human T-cell leukaemia virus type 1 infection, a retroviral disease that is characterized by deep proviral latency that limits active viral transcription and replication, such that viral propagation occurs almost exclusively by mitotic spread during clonal proliferation of infected T cells[13]. On the basis of these considerations, we hypothesized that genome-intact proviral sequences from elite controllers maintain a state of deep, long-lasting latency, possibly owing to chromosomal integration into genomic regions that are not permissive to active viral transcription.

To investigate the chromosomal positions of genome-intact proviral sequences, we used matched integration site and proviral sequencing (MIP-seq)[14] to analyse integration sites together with the corresponding proviral sequences. In brief, proviral DNA was diluted to single-genome levels, amplified by Φ29-catalysed whole-genome amplification and analysed with near-full-length proviral sequencing[14] and integration site analysis using 'integration site loop amplification'[15] or ligation-mediated PCR[16]. These experiments, performed on samples from the eleven elite controllers (EC3–EC13), identified a total of 92 integration sites that corresponded to genome-intact proviral sequences, of which 33 were associated with unique chromosomal locations (Supplementary Table 1). These integration sites of genome-intact proviral sequences were preferentially located in chromosomes 7, 17 and 19, and to a lesser extent in chromosomes 16 and 18 (Fig. 4a and Extended Data Fig. 5a). Consistent with previous studies[14] in which a total of 100 pairs of genome-intact proviral sequences and corresponding integration sites ($n = 73$ genome-intact proviral sequences with unique integration sites) were analysed for long-term ART-treated individuals, proviral species that displayed complete sequence identity shared the same integration sites, which confirms their clonal origin. Notably, upstream HIV-1 long-terminal repeat regions, which are not included in typical FLIP-seq assays[10,12] but that were specifically amplified in sequences from these individuals, also displayed complete sequence identity within analysed clonal proviral sequences (Extended Data Fig. 4).

Notably, integration site analysis revealed that a significantly larger proportion of genome-intact proviral sequences from elite controllers were located in non-genic or pseudogenic regions, relative to genome-intact proviral sequences from long-term ART-treated individuals analysed using the same approach[14] (45% compared with 17.8% of distinct genome-intact proviral sequences, respectively, $P = 0.0051$; 40.2% compared with 13% of all genome-intact proviral sequences, respectively, $P < 0.0001$), and in comparison to previous studies in which integration sites of HIV-1 proviral sequences from ART-treated individuals[15,17] were analysed without distinguishing intact from defective proviruses (Fig. 4b and Extended Data Fig. 5b). Further investigation revealed that the non-genic integration sites of genome-intact proviral sequences from elite controllers were frequently positioned in or surrounded by centromeric satellite or microsatellite DNA (EC3–EC7; Fig. 2a–e), non-coding regions of the human genome that consist of dense heterochromatin 'gene deserts'[18] that are typically disfavoured for HIV-1 integration[19]. Localization of proviral sequences in such centromeric satellite DNA has been associated with deep viral latency in functional viral reactivation studies[20,21] and was extremely rare[22]

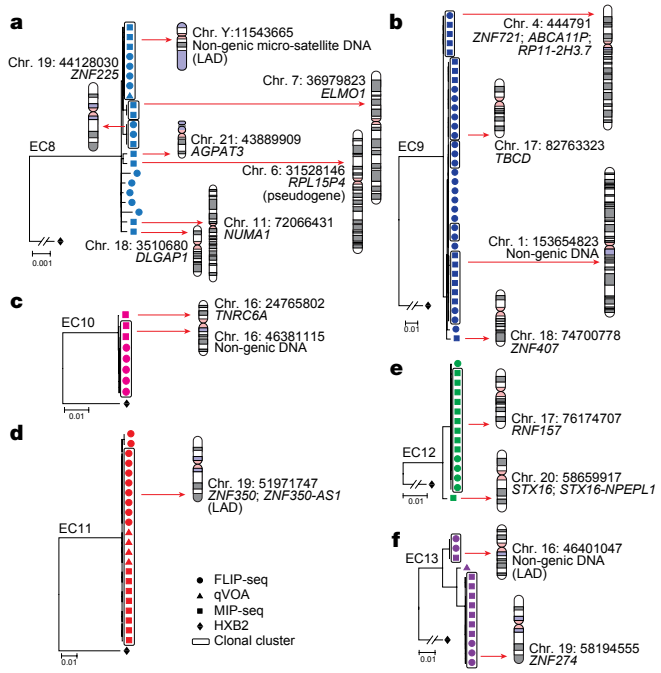


Fig. 3 | Preferential location of genome-intact proviral sequences from elite controllers in genes that encode KRAB-ZNF proteins. a–f, Linear maximum-likelihood phylogenetic trees of genome-intact proviral sequences from the indicated study participants are shown. Coordinates and relative positioning of integration sites are indicated. Other information is as described in the legend of Fig. 2.

or entirely undetectable in previous studies of ART-treated individuals[14]. In our study, the integration of genome-intact proviral sequences into centromeric satellite or microsatellite DNA was observed in a total of 8 unique genome-intact proviral sequences (24% of distinct genome-intact proviral sequences, 20.7% of all genome-intact proviral sequences) and occurred at least once in 5 (EC3–5, EC7 and EC8) (Figs. 2a–c, e, 3a) of the 11 elite controllers analysed. In addition, three integration sites of genome-intact proviral sequences were located in centromeric non-genic DNA surrounded by satellite DNA (EC3 and EC6) (Fig. 2a, d). Notably, as many as six different integration sites of genome-intact proviral sequences were located in or surrounded by centromeric satellite DNA in EC3 (Fig. 2a). In addition to this highly disproportionate overrepresentation of centromeric satellite DNA among integration sites of genome-intact proviral sequences from elite controllers, sequences from EC10 and EC13 contained integrations of clonal genome-intact proviral sequences in a large non-genic region in proximity to non-centromeric microsatellite DNA on chromosome 16 (Fig. 3c, f). Thus, in total, 39.4% of all 33 distinct genome-intact proviral sequences (32.6% of all 92 genome-intact proviral sequences) from elite controllers were located within or in proximity to satellite or microsatellite DNA.

Corresponding to the disproportionate enrichment of non-genic integration sites in elite controllers, we noted that the number of genic integration sites associated with genome-intact proviral sequences was significantly decreased in elite controllers, relative to ART-treated individuals[14]. These genic integration sites were almost exclusively located in introns of genes that, in comparison to long-term ART-treated individuals, showed weaker transcriptional activity (Extended Data Fig. 7a) and displayed an opposite orientation relative to the host gene, in which the proviral sequence was contained, in approximately 60% of all genic integration sites analysed (Extended Data Fig. 7b, c). Genes that encode members of the zinc-finger protein (ZNF) family and, in particular, Krüppel-associated box domain-containing ZNF (KRAB-ZNF) genes[23] accounted for 33% of all 18 genes that contained distinct genome-intact proviral sequences in elite controllers (corresponding to 49% of all 55 genic integration events of genome-intact proviral sequences), a notable enrichment relative to ART-treated individuals (Fig. 4c and Extended Data Fig. 5c). Clonal genome-intact proviral sequences were frequently integrated into KRAB-ZNF genes located in defined regions of chromosome 19[24] that display highly distinct chromatin features. In particular, these regions are extensively occupied by the heterochromatin proteins CBX1 and SUV39H1[25] and also show a strong enrichment for repressive chromatin marks that cover the lengths of ZNF genes but selectively spare the corresponding host transcriptional start sites[25]. Notably, a previous computational, genome-wide analysis of chromatin states based on the combinatorial evaluation of multiple different chromatin marks in their respective spatial context revealed that repetitive satellite DNA and ZNF genes share a common, highly distinct chromatin state (referred to as 'ZNF genes and repeats')[26]. When combined, genome-intact proviral sequences located either in satellite DNA or in ZNF genes represented more than 45% of all 33 independent genome-intact proviral sequences and more than 60% of all 92 genome-intact proviral sequences in elite controllers, proportions that were significantly increased relative to ART-treated individuals (Fig. 4d and Extended Data Fig. 5d).

To analyse the positioning of proviral integration sites relative to active transcription units in host DNA, we performed RNA-sequencing-based gene-expression profiling in autologous total $CD4^+$ T cells, as well as autologous central memory and effector memory $CD4^+$ T cell subsets, which contain the majority of the viral reservoir cells in peripheral blood[27]. These experiments showed a significantly increased chromosomal distance between the integration sites of genome-intact proviral sequences and the most proximal host transcriptional start sites in elite controllers, relative to long-term ART-treated individuals[14] (Fig. 4e). Simultaneously, we calculated the chromosomal distance between the coordinates of integration sites of genome-intact proviral sequences and accessible chromatin, as determined by genome-wide assays for transposase-accessible chromatin using sequencing (ATAC-seq) performed in autologous $CD4^+$ T cells. Although integration sites in satellite and microsatellite DNA were excluded from this analysis (and from the subsequent analysis using chromatin immunoprecipitation followed by sequencing (ChIP–seq), high-throughput chromatin conformation capture sequencing (Hi-C-seq) and methylation-sequencing data; see below) due to the reduced ability to map next-generation sequencing reads onto repetitive genomic DNA regions[28], we noted that integration sites of genome-intact proviruses from elite controllers were located at significantly increased distances from accessible chromatin, compared to those from ART-treated individuals[14] (Fig. 4f). These differences were observed when clonal sequences were counted only once (Fig. 4e, f) but were also notable when all clonal sequences were considered individually (Extended Data Fig. 5e, f).

In a subsequent analysis, we calculated the number of DNA reads associated with defined epigenetic histone marks in proximity to viral integration sites using ChIP–seq data from primary memory $CD4^+$ T cells available from the ROADMAP Epigenomics Project[26]. In comparison to ART-treated individuals[14], this analysis revealed a marked enrichment of the repressive histone feature H3K9me3 (on chromosomes 7 and 19) and/or a de-enrichment of the activating chromatin feature H3K4me1 (on chromosomes 17 and 19) at integration sites of genome-intact proviral sequences from elite controllers (Fig. 4g); a trend for differential expression of additional activating and inhibitory chromatin modifications in proximity to integration sites of genome-intact proviral sequences from elite controllers and ART-treated individuals was also noted (Extended Data Fig. 6a–d). Furthermore, an alignment of the coordinates of integration sites to three-dimensional chromosomal contact data generated by Hi-C-seq[29] demonstrated a significantly increased proportion of genome-intact proviral sequences from elite controllers located in compartment B, which mostly contains closed

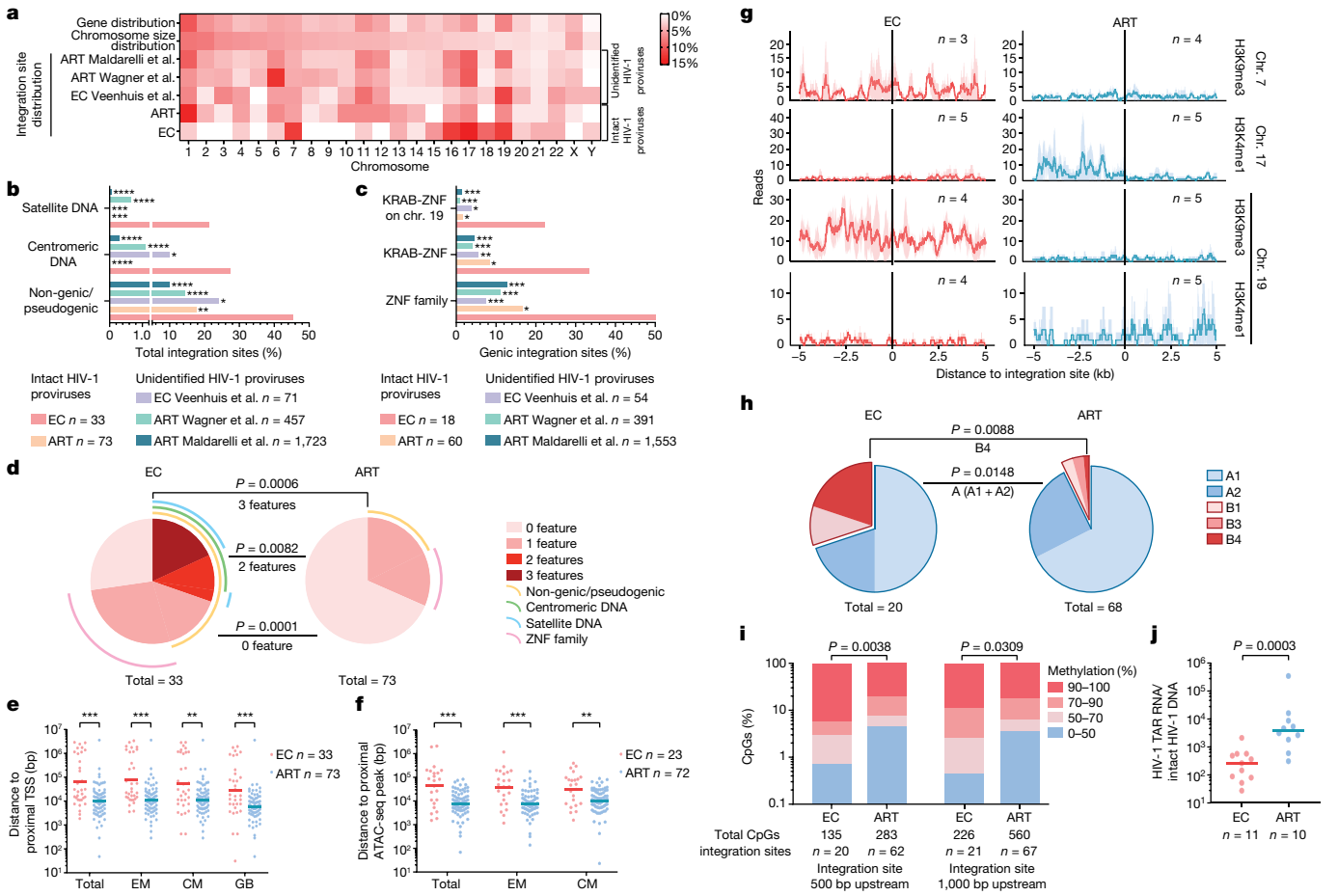


Fig. 4 | Distinct genomic and epigenetic features of integration sites of genome-intact proviral sequences from elite controllers. a, Relative proportion of proviral integration sites of genome-intact proviral sequences in each chromosome. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of genome-intact proviral sequences located in the indicated genomic regions. **a**–**c**, Data from genome-intact proviral sequences in ART-treated individuals[14] and from unselected (intact and defective) proviral sequences from elite controllers (Veenhuis et al., ref. [9]) and ART-treated individuals (Wagner et al., ref. [15] and Maldarelli et al., ref. [17]) are shown as references. **d**, SPICE diagrams show the proportions of genome-intact proviral sequences with the indicated integration site features in elite controllers and ART-treated individuals. **e**, **f**, Chromosomal distance between integration sites of genome-intact proviral sequences and the most proximal transcriptional start sites (TSS) in autologous total, effector memory or central memory CD4$^+$ T cells or from the Genome Browser (GB) (**e**), or to the most proximal ATAC-seq peaks (**f**) in autologous total, effector memory and central memory CD4$^+$ T cells. Horizontal lines show the geometric mean. **g**, Numbers of DNA-sequencing reads associated with activating (H3K4me1) or repressive (H3K9me3) histone protein modifications in proximity to integration sites from elite controllers and long-term ART-treated individuals[14]. Median and confidence intervals (one standard deviation) of ChIP–seq data from primary memory CD4$^+$ T cells included in the ROADMAP repository[26] are shown. **h**, Proportions of genome-intact proviral sequences located in structural compartments A and B (and associated sub-compartments), as determined by Hi-C-seq data[29]. Integration sites in regions not covered in a previous study[29] were excluded. **i**, Numbers of cytosine residues with indicated levels of methylation (derived from CD4$^+$ T cells in the iMethyl database[31]) in proximity (500 or 1,000 bp upstream of the 5′ long-terminal repeat (LTR) host–viral junction) to integration sites from elite controllers and ART-treated individuals. **j**, Frequencies of HIV-1 RNA transcripts in PBMCs from elite controllers and ART-treated individuals, normalized to the corresponding number of genome-intact proviral sequences determined by FLIP-seq. **a**–**i**, Clonal sequences were only counted once. **f**–**i**, Sequences in genomic regions included in the ENCODE blacklist[28] were excluded. ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$; data were analysed using two-sided Fisher's exact tests (**b**–**d**, **h**), two-sided Mann–Whitney *U*-tests (**e**, **f**, **j**) or two-tailed χ^2 test (**i**); **b**, **c**, **e**, **f**, **i**, FDR-adjusted *P* values are shown; **d**, **h**, **j**, nominal *P* values are shown. All comparisons were made between elite controllers and reference groups.

chromatin. This effect was particularly obvious for integration sites in KRAB-ZNF genes on chromosome 19 in elite controllers, which were all located in subcompartment B4 (Fig. 4h and Extended Data Fig. 5g). This very small compartment (which accounts for approximately 0.3% of the human genome) is known to contain dense heterochromatin marks[29] and represents a highly atypical location of a chromosomal integration site for HIV-1 in non-controller individuals[14]. A highly increased frequency of genome-intact proviral sequences from elite controllers in compartment B was also noted when Hi-C-seq data from Jurkat cells[30] were used for alignment (Extended Data Fig. 6e, f).

Taking advantage of previously published genome-wide bisulfite sequencing data of CD4$^+$ T cells[31], we observed that the frequency of hypermethylated (more than 90% methylation) cytosine residues was significantly higher in proximity to genome-intact proviral sequences from elite controllers, relative to integration sites of genome-intact proviral sequences from long-term ART-treated individuals[14] (Fig. 4i). These data suggest that chromosomal regions that are more susceptible to DNA methyltransferases represent preferential sites for the long-term persistence of genome-intact proviral sequences in elite controllers, arguably because the integration into hypermethylated genomic DNA might facilitate deep latency of genome-intact proviral sequences and protect against immune-cell targeting. Given that closely neighbouring cytosine residues are likely to share the same methylation status[32], these results raise the possibility that HIV-1

promoter methylation, which has previously been shown to induce proviral HIV-1 silencing in in vitro assays[33], may contribute to durable transcriptional repression of genome-intact proviral sequences from elite controllers. The frequencies of genome-intact proviral sequences located in lamina-associated domains—genomic regions that interact with the inner nuclear membrane, mostly contain closed chromatin and represent a rare target for HIV-1 integration[34]—were not significantly different between genome-intact proviral sequences from elite controllers and ART-treated individuals when clonal sequences were counted only once; however, a significant enrichment of genome-intact proviral sequences from elite controllers in lamina-associated domains was noted when clonal genome-intact proviral sequences were counted as independent proviruses (Extended Data Fig. 7d, e).

Given that non-coding centromeric satellite DNA is a highly disfavoured target site for HIV-1 integration[19], the disproportionately increased number of integration sites in satellite DNA described here is a remarkable feature of elite controllers. Notably, elite controllers expressed normal mRNA levels of *LEDGF* (also known as *PSIP1* or *p75*) and *CPSF6* (Extended Data Fig. 7f), host factors that interact directly with HIV-1 proteins to bias HIV-1 integration site selection to active transcription units[35,36]. Although protein levels of these molecules were not assessed, these results suggest that there is no increased susceptibility of centromeric satellite DNA to HIV-1 integration in elite controllers. To further address this, we infected $CD4^+$ T cells from $n = 12$ elite controllers from our study cohort and $n = 9$ HIV-1-negative healthy individuals with a GFP-encoding HIV-1 construct, followed by sorting of GFP^+ and GFP^- $CD4^+$ T cells and subsequent integration site analysis. These experiments, in which more than 120,000 independent HIV-1 integration coordinates were obtained, showed that integration sites in satellite DNA accounted for extremely low proportions of all integration events (0.04–0.06% in GFP^+ and 0.11–0.12% in GFP^- $CD4^+$ T cells), irrespective of the analysed study cohort (Extended Data Fig. 8a, b and Supplementary Table 2). Moreover, there was no evidence for preferential targeting of non-genic chromosomal regions or genes that encode KRAB-ZNF proteins in $CD4^+$ T cells from elite controllers that were infected in vitro (Extended Data Fig. 8b, c).

In conclusion, this work identifies a markedly distinct reservoir landscape of intact proviral sequences in PBMCs from individuals with durable natural control of HIV-1, characterized by features of integration sites that are highly suggestive of deep latency. For additional functional validation of this conclusion, we analysed the frequency of cell-associated HIV-1 RNA transcripts in elite controllers and ART-treated individuals; these additional experiments demonstrated that the number of cell-associated HIV-1 RNA copies, normalized to the corresponding number of genome-intact proviral sequences, was significantly lower in elite controllers (Fig. 4j). As such, elite controllers seem to exemplify attributes of a 'block and lock' mechanism[37] of viral control, which is defined by silencing of proviral gene expression through chromosomal integration into repressive chromatin locations[38]. We propose that the distinct reservoir configuration in elite controllers is not related to altered preferences for integration site locations during acute HIV-1 infection in elite controllers, but instead represents the result of cell-mediated immune selection forces that preferentially eliminate proviral sequences that are more permissive to viral transcription, in a process that we suggest referring to as the 'autologous shock and kill' mechanism. By contrast, less transcriptionally active proviral sequences with features of deep latency, leading to lower vulnerability to immune recognition, seem to persist long-term. In very rare cases, such as in EC1 and EC2, such selection forces may have accomplished near-complete clearance of all genome-intact proviral sequences, raising the possibility that a sterilizing cure of HIV-1 infection can, at least in principle, spontaneously occur through natural, immune-mediated mechanisms. Future studies will be necessary to determine whether signs of immune-mediated selection pressure on viral reservoir cells are also found in genome-intact proviral sequences from lymphoid tissues, which contain the majority of viral reservoir cells[39].

Although our data strongly suggest that deep latency has a role in maintaining spontaneous, drug-free control of HIV-1 in some elite controllers, deep viral latency is not completely permanent or irreversible, as reflected by our ability to retrieve replication-competent virus from elite controllers in in vitro qVOAs. However, in vitro qVOAs with maximum stimuli are unlikely to adequately reflect the susceptibility to viral reactivation in vivo; indeed, in vitro viral outgrowth may largely be a stochastic process[12,40], and may occur independently of molecular pathways that fine-tune the outgrowth behaviour of the virus in vivo. Nevertheless, it is likely that deep viral latency in elite controllers is a dynamic process, and that occasional bursts of viral transcription may occur despite genomic and epigenetic features of integration sites restricting viral gene expression. In fact, a proviral landscape with low permissiveness to viral reactivation stimuli may expose the immune system to a tailored viral antigen dose that can maintain a highly functional antiviral T cell response, a hallmark of antiviral immunity in elite controllers[5], without supporting high-level viral replication promoting cytotoxic T cell exhaustion. Therefore, a reciprocal equilibrium between a weakly inducible viral reservoir and an efficient HIV-1-specific $CD8^+$ T cell response may represent the cornerstone of natural HIV-1 immune control. Given that evidence for selection of genome-intact proviral sequences with features of deeper latency was also observed in long-term ART-treated individuals, albeit to a weaker degree[14], it is hoped that future longitudinal evaluations will be informative for designing strategies to induce long-term drug-free remission of HIV-1 infection in larger populations of individuals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2651-8.

1. Sáez-Cirión, A. & Pancino, G. HIV controllers: a genetically determined or inducible phenotype? *Immunol. Rev.* **254**, 281–294 (2013).
2. Yukl, S. A. et al. Challenges in detecting HIV persistence during potentially curative interventions: a study of the Berlin patient. *PLoS Pathog.* **9**, e1003347 (2013).
3. Gupta, R. K. et al. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature* **568**, 244–248 (2019).
4. McLaren, P. J. & Carrington, M. The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.* **16**, 577–583 (2015).
5. Migueles, S. A. et al. Lytic granule loading of $CD8^+$ T cells is required for HIV-infected cell elimination associated with immune control. *Immunity* **29**, 1009–1021 (2008).
6. Gaiha, G. D. et al. Structural topology defines protective CD8+ T cell epitopes in the HIV proteome. *Science* **364**, 480–484 (2019).
7. Migueles, S. A. & Connors, M. Success and failure of the cellular immune response against HIV-1. *Nat. Immunol.* **16**, 563–570 (2015).
8. Blankson, J. N. et al. Isolation and characterization of replication-competent human immunodeficiency virus type 1 from a subset of elite suppressors. *J. Virol.* **81**, 2508–2518 (2007).
9. Veenhuis, R. T. et al. Long-term remission despite clonal expansion of replication-competent HIV-1 isolates. *JCI Insight* **3**, e122795 (2018).
10. Lee, G. Q. et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 $CD4^+$ T cells. *J. Clin. Invest.* **127**, 2689–2696 (2017).
11. Popper, K. *Die Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft* (Springer, 1935).
12. Ho, Y. C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
13. Melamed, A. et al. Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. *PLoS Pathog.* **9**, e1003271 (2013).
14. Einkauf, K. B. et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J. Clin. Invest.* **129**, 988–998 (2019).
15. Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
16. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
17. Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).

18. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res*. **26**, 115–138 (2018).
19. Carteau, S., Hoffmann, C. & Bushman, F. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J. Virol*. **72**, 4005–4014 (1998).
20. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J*. **22**, 1868–1877 (2003).
21. Lewinski, M. K. et al. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol*. **79**, 6610–6619 (2005).
22. Schröder, A. R. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
23. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017).
24. Lukic, S., Nicolas, J. C. & Levine, A. J. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death Differ*. **21**, 381–387 (2014).
25. Vogel, M. J. et al. Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res*. **16**, 1493–1504 (2006).
26. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
27. Chomont, N. et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Med*. **15**, 893–900 (2009).
28. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep*. **9**, 9354 (2019).
29. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
30. Lucic, B. et al. Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nat. Commun*. **10**, 4059 (2019).
31. Komaki, S. et al. iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation. *Hum. Genome Var*. **5**, 18008 (2018).
32. Affinito, O. et al. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics* **112**, 144–150 (2020).
33. Kauder, S. E., Bosque, A., Lindqvist, A., Planelles, V. & Verdin, E. Epigenetic regulation of HIV-1 latency by cytosine methylation. *PLoS Pathog*. **5**, e1000495 (2009).
34. Marini, B. et al. Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**, 227–231 (2015).
35. Ciuffi, A. et al. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med*. **11**, 1287–1289 (2005).
36. Achuthan, V. et al. Capsid–CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24**, 392–404 (2018).
37. Debyser, Z., Vansant, G., Bruggemans, A., Janssens, J. & Christ, F. Insight in HIV integration site selection provides a block-and-lock strategy for a functional cure of HIV infection. *Viruses* **11**, 12 (2018).
38. Battivelli, E. et al. Distinct chromatin functional states correlate with HIV latency reactivation in infected primary $CD4^+$ T cells. *eLife* **7**, e34655 (2018).
39. Estes, J. D. et al. Defining total-body AIDS-virus burden with implications for curative strategies. *Nat. Med*. **23**, 1271–1276 (2017).
40. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants

HIV-1-infected study participants were recruited at the Massachusetts General Hospital (MGH), the Brigham and Women's Hospital (BWH) and at the University of California, San Francisco (UCSF) at the Zuckerberg San Francisco General Hospital. PBMCs and tissue samples were obtained according to protocols approved by the respective Institutional Review Boards. Clinical and demographical characteristics of study participants are summarized in Extended Data Table 1.

Droplet digital PCR

DNA was extracted from PBMCs or $CD4^+$ T cells isolated from total PBMCs (CD4 T Cell Isolation Kit, Miltenyi Biotec, 130-096-533) using commercial kits (Qiagen, DNeasy, 69504). We amplified total HIV-1 DNA using droplet digital PCR (ddPCR; Bio-Rad), using primers and probes that have previously been described[10] (127-bp 5′LTR–*gag* amplicon; HXB2 coordinates 684–810). PCR was performed using the following program: 95 °C for 10 min, 45 cycles of 94 °C for 30 s and 60 °C for 1 min, 72 °C for 1 min. The droplets were subsequently read by the QX200 droplet reader and data were analysed using QuantaSoft software (Bio-Rad).

Whole-genome amplification

Extracted DNA was diluted to single viral genome levels according to ddPCR results, so that 1 provirus was present in approximately 20–30% of wells. Subsequently, DNA in each well was subjected to multiple displacement amplification (MDA) with Φ29 polymerase (Qiagen, REPLI-g Single Cell Kit, 150345), as per the manufacturer's protocol. Following this unbiased whole-genome amplification step[41], DNA from each well was split and separately subjected to viral sequencing and integration site analysis, as described below. If necessary, a second-round multiple displacement amplification reaction was performed to increase the amount of available DNA.

HIV-1 near-full-genome sequencing

DNA resulting from whole-genome amplification reactions was subjected to near-full-length HIV-1 genome amplification using a one-amplicon and/or non-multiplexed five-amplicon approach, as previously described[14]. PCR products were visualized by agarose gel electrophoresis (Quantify One and ChemiDoc MP Image Lab, Bio-Rad). All near-full-length and/or five-amplicon-positive amplicons were subjected to Illumina MiSeq sequencing at the MGH DNA Core Facility. Resulting short reads were de novo assembled using Ultracycler v.1.0 and aligned to HXB2 to identify large deleterious deletions (<8,000 bp of the amplicon aligned to HXB2), out-of-frame indels, premature/lethal stop codons, internal inversions or packaging signal deletions (≥15 bp insertions and/or deletions relative to HXB2), using an automated in-house pipeline written in Python programming language (https://github.com/BWH-Lichterfeld-Lab/Intactness-Pipeline), consistent with previous studies[10,42,43,44]. The presence or absence of APOBEC-3G/3F-associated hypermutations was determined using the Los Alamos National Laboratory (LANL) HIV Sequence Database Hypermut 2.0[45] program. Viral sequences that lacked all mutations listed above were classified as 'genome-intact' sequences. Sequence alignments were performed using MUSCLE[46]. Phylogenetic distances between sequences were examined using maximum-likelihood trees in MEGA (https://www.megasoftware.net/) and MAFFT (https://mafft.cbrc.jp/alignment/software), and visualized using Highlighter plots (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html). Viral sequences were considered clonal if they had completely identical consensus sequences; single-nucleotide variations in primer-binding sites were not considered for clonality analysis. Clades of intact HIV-1 proviral sequences were determined using the LANL HIV-1 Sequence Database Recombinant Identification Program[47]. Within intact HIV-1 clade B sequences, the proportions of optimal CTL epitopes (restricted by autologous HLA class I alleles) that match the clade B consensus sequence and CTL escape variants restricted by selected HLA class I alleles and supertypes described in the LANL HIV Immunology Database (https://www.hiv.lanl.gov/content/index) were determined.

Integration site analysis

Integration sites associated with each proviral sequence were obtained using integration site loop amplification, as previously described[15], or by ligation-mediated PCR[16] (Lenti-X Integration Site Analysis Kit (Takara Bio, 631263)); DNA produced by whole-genome amplification was used as template. For selected clonal sequences, viral–host junction regions were also amplified using primers that anneal upstream of the integration site in host DNA and downstream of the integration site in viral DNA. Resulting PCR products were subjected to next-generation sequencing using Illumina MiSeq. MiSeq paired-end FASTQ files were demultiplexed; small reads (142 bp) were then aligned simultaneously to the human reference genome GRCh38 and HIV-1 reference genome HXB2 using bwa-mem[48]. Biocomputational identification of integration sites was performed according to previously described procedures[15,49]. In brief, chimeric reads containing both human and HIV-1 sequences were evaluated for mapping quality based on (1) HIV-1 coordinates mapping to the terminal nucleotides of the viral genome; (2) absolute counts of chimeric reads; and (3) depth of sequencing coverage in the host genome adjacent to the viral integration site. The final list of integration sites and corresponding chromosomal annotations was obtained using Ensembl (v.86, http://www.ensembl.org/index.html), the UCSC Genome Browser (http://www.genome.ucsc.edu/) and GENCODE (v.29, https://www.gencodegenes.org/). Repetitive genomic sequences containing HIV-1 integration sites were identified using RepeatMasker (http://www.repeatmasker.org/).

Cell sorting and flow cytometry

PBMCs were stained with monoclonal antibodies against CD4 (1:50, clone RPA-T4, Biolegend, 300518), CD3 (1:50, clone OKT3, Biolegend, 317332), CD45RO (1:40, clone UCHL1, Biolegend, 304236) and CCR7 (1:40, clone G043H7, Biolegend, 353216). Afterwards, cells were washed and $CD45RO^+CCR7^+$ (central memory), $CD45RO^+CCR7^-$ (effector memory) and $CD3^+CD4^+$ (total) $CD4^+$ T cells were sorted in a specifically designated biosafety cabinet (Baker Hood), using a FACS Aria cell sorter (BD Biosciences) at 70 pounds per square inch. Cell sorting was performed by the Ragon Institute Imaging Core Facility at MGH and resulted in isolation of lymphocytes with the defined phenotypic characteristics of >95% purity. Data were analysed using FlowJo software (Treestar).

RNA-seq

Total RNA was extracted from sorted $CD4^+$ T cell populations using a PicoPure RNA Isolation Kit (Applied Biosystems, KIT0204). RNA-seq libraries were generated as previously described[50]. In brief, whole-transcriptome amplification and tagmentation-based library preparation was performed using SMART-seq2, followed by sequencing on a NextSeq 500 Instrument (Illumina). The quantification of transcript abundance was conducted using RSEM software (v1.2.22) supported by STAR aligner software (STAR 2.5.1b) and aligned to the GRCh38 human genome. Transcripts per million values were then normalized among all samples using the upper-quantile-normalization method.

ATAC-seq

A previously described protocol with some modifications[51,52] was used. In brief, 20,000 sorted cells were centrifuged at 1,500 rpm for 10 min

at 4 °C in a pre-cooled fixed-angle centrifuge. All of the supernatant was removed and a modified transposase mixture (including 25 µl of 2× TD buffer, 1.5 µl of TDE1, 0.5 µl of 1% digitonin, 16.5 µl of PBS, 6.5 µl of nuclease-free water) was added to the cells and incubated in a heat block at 37 °C for 30 min. Transposed DNA was purified using a ChIP DNA Clean & Concentrator Kit (Zymo Research, D5205) and eluted DNA fragments were used to amplify libraries. The libraries were quantified using an Agilent Bioanalyzer 2100 and the Qubit dsDNA High Sensitivity Assay Kit. All Fast-ATAC libraries were sequenced using paired-end, single-index sequencing on a NextSeq 500/550 instrument with v.2.5 Kits (75 cycles). The quality of reads was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk). Low-quality DNA end fragments and sequencing adapters were trimmed using Trimmomatic (http://www.usadellab.org). Sequencing reads were then aligned to the human reference genome GRCh38 using a short-read aligner (Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) with the non-default parameters 'X2000', 'non-mixed' and 'non-discordant'. Reads from mitochondrial DNA were removed using Samtools (http://www.htslib.org). Peak calls were made using MACS2 with the callpeak command (https://pypi.python.org/pypi/MACS2), with a threshold for peak calling set to FDR-adjusted $P < 0.05$.

qVOAs

$CD4^+$ cells were isolated from PBMCs using the EasySep Human CD4 Positive Selection Kit II (STEMCELL Technologies 17852). Cells were plated in limiting dilutions based on the intact provirus reservoir size determined through FLIP-seq. Irradiated feeder PBMCs were added at 1×10^5 cells per well. Cells were activated with 1 µg/ml PHA for 4 days, which was subsequently washed away and 10,000 MOLT-4 $CCR5^+$ cells (NIH AIDS Reagent Program, 4984) were added to propagate infection. On the 13th and 20th days, culture supernatants from each well were individually incubated with 10,000 TZM-bl cells (NIH AIDS Reagent Program, 8129) to drive *Tat*-dependent luciferase production. On the 15th and 22nd days, TZM-bl cells were lysed, and luciferase activity was measured using Britelite Plus (PerkinElmer, 6066761). Luciferase-positive wells were defined as having signal levels that were >3-fold higher than negative controls. Cells from positive wells were then collected and plated into bottom compartments of Transwell tissue-culture inserts (Costar 6.5 mm Transwells, 0.4-µm pore polyester membrane inserts, STEMCELL, 38024), while 1×10^6 MOLT-4 cells were placed in top compartments. After five additional days of culture, MOLT-4 cells from the upper wells were collected and subjected to FLIP-seq. Large-scale quantitative viral outgrowth measurements on cells from patient EC2 were performed by a similar standard method[53] with a p24 ELISA assay used to detect outgrowth.

Intact proviral DNA assays

The intact proviral DNA assay (IPDA) uses ddPCR to quantify proviruses that lack overt fatal defects, especially large deletions and hypermutations, and was performed as previously described[54].

In vitro-infection assays

$CD4^+$ T cells were stimulated in RPMI medium supplemented with 10% fetal calf serum, recombinant IL-2 (50 U/ml), and an anti-CD3/CD8 bispecific antibody (0.5 µg/µl, NIH AIDS Reagent Program, 12277). Cells were infected on day 5 with a GFP-encoding NL4-3 construct with a BAL-derived R5-tropic envelope[55] at a multiplicity of infection (MOI) of 0.1 for 4 h at 37 °C. After two washes, cells were resuspended in medium and plated at 5×10^5 cells per well in a 24-well plate. On day 5, GFP^+ and GFP^- $CD4^+$ T cells were sorted. Cells were processed for DNA extraction and integration site analysis using ligation-mediated PCR according to a previously described protocol[49].

Analysis of cell-associated HIV-1 RNA

Total cell-associated RNA and DNA was extracted in parallel from the same PBMC sample, using the GenElute RNA/DNA/Protein Purification Plus Kit (Sigma RDP300) according to the manufacturer's protocol. RNA was reverse-transcribed into cDNA using a polyadenylation-RT reaction[56] to efficiently detect HIV-1 RNA transcripts, followed by ddPCR-based amplification with primers and probes that span the HIV-1 trans-activation response (tar) region, as described previously[56]. Simultaneously, cell-associated DNA was subjected to ddPCR-based amplification of the *RPP30* gene to determine cell counts in PBMC samples, using probes and primers described previously[57]. Cell-associated HIV-1 RNA copies per million PBMCs were normalized to the corresponding number of intact proviruses per million PBMCs (determined by FLIP-seq).

Statistics

Data are shown as pie charts, bar charts, scatter plots with individual values or heat maps. Differences were tested for statistical significance using Mann–Whitney *U*-tests (two-tailed), Fisher's exact tests (two-tailed) or χ^2 tests (two-tailed), as appropriate. $P < 0.05$ was considered significant, FDR correction was performed using the Benjamini–Hochberg method[58]. Analyses were performed using Prism (GraphPad Software), SPICE[59] and R (R Foundation for Statistical Computing)[60].

Study approval

Study participants gave written informed consent to participate in accordance with the Declaration of Helsinki. The study was approved by the Institutional Review Boards of MGH, BWH and UCSF.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-seq and ATAC-seq data have been deposited in the NCBI GEO (accession number GSE144334). Owing to study participant confidentiality concerns, full-length viral sequencing data cannot be publicly released, but will be made available to investigators upon reasonable request and after signing a coded tissue agreement. The Los Alamos HIV Sequence Database Hypermut 2.0 and the Los Alamos HIV Immunology Database 2.0 are available at https://www.hiv.lanl.gov/content/index. The iMethyl database is available at http://imethyl.iwate-megabank.org. ROADMAP epigenomic data are available at http://www.roadmapepigenomics.org.

41. Burtt, N. P. Whole-genome amplification using Φ29 DNA polymerase. *Cold Spring Harb. Protoc.* **2011**, pdb.prot5552 (2011).
42. Lee, G. Q. et al. HIV-1 DNA sequence diversity and evolution during acute subtype C infection. *Nat. Commun.* **10**, 2737 (2019).
43. Hiener, B. et al. Identification of genetically intact HIV-1 proviruses in specific $CD4^+$ T cells from effectively treated participants. *Cell Rep.* **21**, 813–822 (2017).
44. Pinzone, M. R. et al. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat. Commun.* **10**, 728 (2019).
45. Rose, P. P. & Korber, B. T. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. *Bioinformatics* **16**, 400–401 (2000).
46. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
47. Siepel, A. C., Halpern, A. L., Macken, C. & Korber, B. T. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* **11**, 1413–1416 (1995).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Serrao, E., Cherepanov, P. & Engelman, A. N. Amplification, next-generation sequencing, and genomic DNA mapping of retroviral integration sites. *J. Vis. Exp.* **109**, e53840 (2016).
50. Trombetta, J. J. et al. Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1–4.22.17 (2014).
51. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
52. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
53. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting $CD4^+$ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).

54. Bruner, K. M. et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019).
55. Chen, H. et al. CD4$^+$ T cells from elite controllers resist HIV-1 infection by selective upregulation of p21. *J. Clin. Invest.* **121**, 1549–1560 (2011).
56. Yukl, S. A. et al. HIV latency in isolated patient CD4$^+$ T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Sci. Transl. Med.* **10**, eaap9927 (2018).
57. Kuo, H. H. et al. Anti-apoptotic protein BIRC5 maintains survival of HIV-1-infected CD4$^+$ T cells. *Immunity* **48**, 1183–1194 (2018).
58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
59. Roederer, M., Nozzi, J. L. & Nason, M. C. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry* **79A**, 167–174 (2011).
60. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/ (R Foundation for Statistical Computing, 2019).
61. Robson, M. I. et al. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* **27**, 1126–1138 (2017).

Acknowledgements X.G.Y. is supported by NIH grants HL134539, AI116228, AI078799, DA047034 and the Bill and Melinda Gates Foundation (INV-002703). M.L. is supported by NIH grants AI098487, AI135940, AI114235, AI117841, AI120008 and DK120387. M.L. and X.G.Y. are Associated Members of the BEAT-HIV Martin Delaney Collaboratory (UM1 AI126620). A.N.E. is supported by NIH grant AI052014. Support was also provided by the Harvard University (HU) and University of California at San Francisco (UCSF)/Gladstone Institute for HIV Cure Research Centers for AIDS Research (P30 AI060354 and P30 AI027763, respectively), which are supported by the following institutes and centers that are co-funded by and associated with the US National Institutes of Health: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH, NIA, FIC and OAR, and by HU CFAR Developmental Awards (S.H.). We thank the MGH DNA core facility. R.F.S. and J.M.S. are supported by the NIH Martin Delaney I4C (UM1 AI126603), BEAT-HIV (UM1 AI126620) and the Delaney AIDS Research Enterprise (DARE; UM1 AI126611) Collaboratories and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation (OPP1115715). Additional support for the SCOPE cohort was provided by DARE (AI096109 and AI127966) and the amfAR Institute for HIV Cure Research (amfAR 109301). G.M.L. is supported by NSF grant 1738428 and NIH grant R44AI124996. The International HIV Controller Cohort is supported by the Bill and Melinda Gates Foundation (OPP1066973), the Ragon Institute of MGH, MIT and Harvard, the NIH (R37 AI067073 to B.D.W.) and the Mark and Lisa Schwartz Family Foundation. This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract no. HHSN261200800001E. This research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research and Intramural Programs of NIDCR, NIH. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Author contributions C.J., X.L., K.B.E., J.M.C., B.R., K.C. and J.E.B. performed whole-genome amplification and HIV-1 sequencing; C.J. and K.B.E. analysed integration sites in cells infected in vivo; E.S. and A.N.E. analysed integration sites in cells infected in vitro; S.H. and X.S. performed RNA-seq; X.S. carried out ATAC-seq; C.J., X.L., K.B.E., J.E.B. and M.O. analysed HIV-1 RNA transcripts; C.G. performed bioinformatics analysis; J.M.C., S.M.Y.C., L.N.B., S.E.S., J.A.V., R.F.S. and J.M.S. carried out qVOAs; M.J.P., R.H., M.S., J.M., P.D.B., T.W.C., S.G.D. and B.D.W. contributed PBMCs and tissue samples; M.C. carried out HLA class I typing; G.M.L., R.F.S. and J.M.S. performed IPDA; C.J., X.L., C.G., M.L. and X.G.Y. carried out data interpretation, analysis and presentation; C.J., X.L., C.G., M.L. and X.G.Y. prepared and wrote the manuscript; C.G., X.S., K.B.E., R.H., A.N.E., M.C., S.G.D., R.F.S. and B.D.W. critically reviewed and edited the manuscript; M.L. and X.G.Y. conceived the research idea and concept and supervised the study.

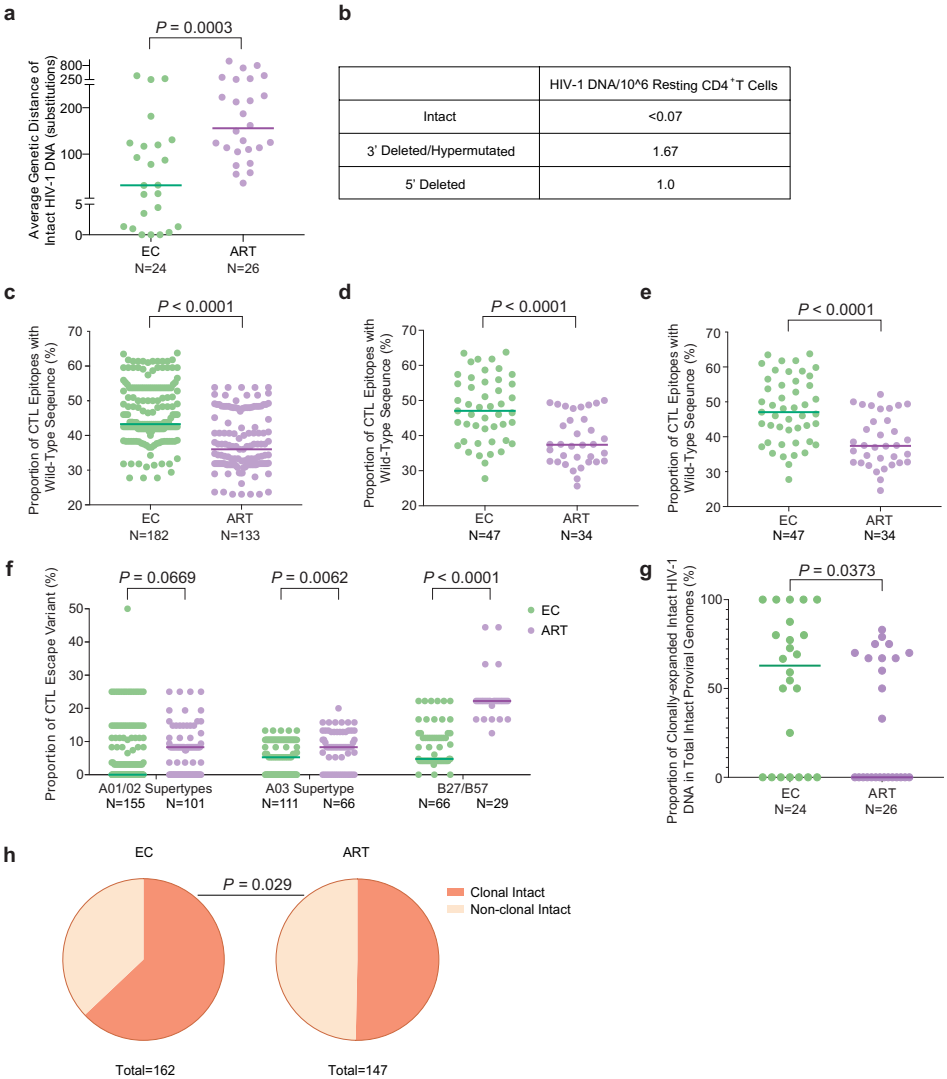
Competing interests A.N.E. has received fees from ViiV Healthcare within the past year for work unrelated to this project. All other authors declare no competing interests.

Additional information

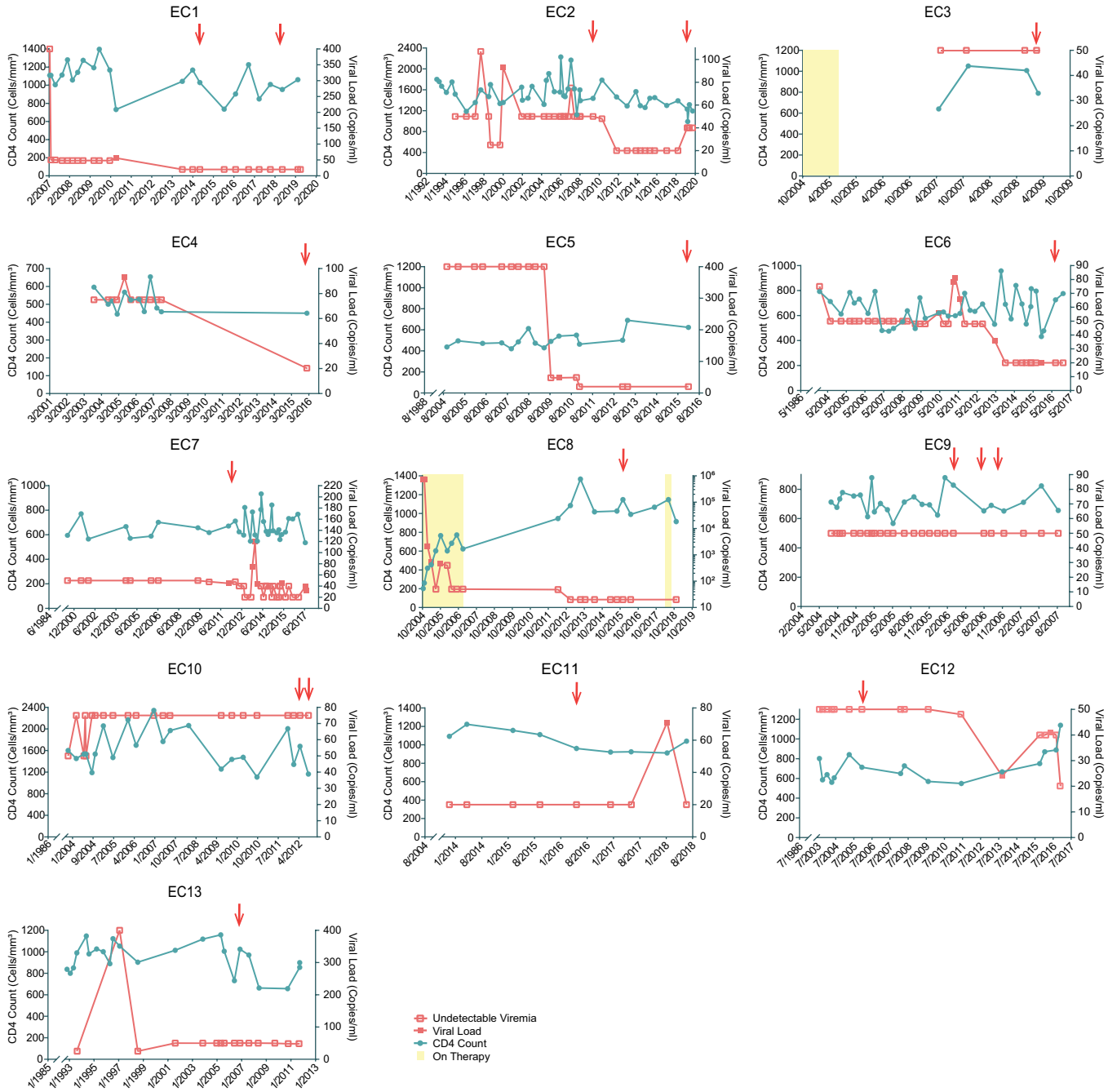
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2651-8.

Correspondence and requests for materials should be addressed to X.G.Y.

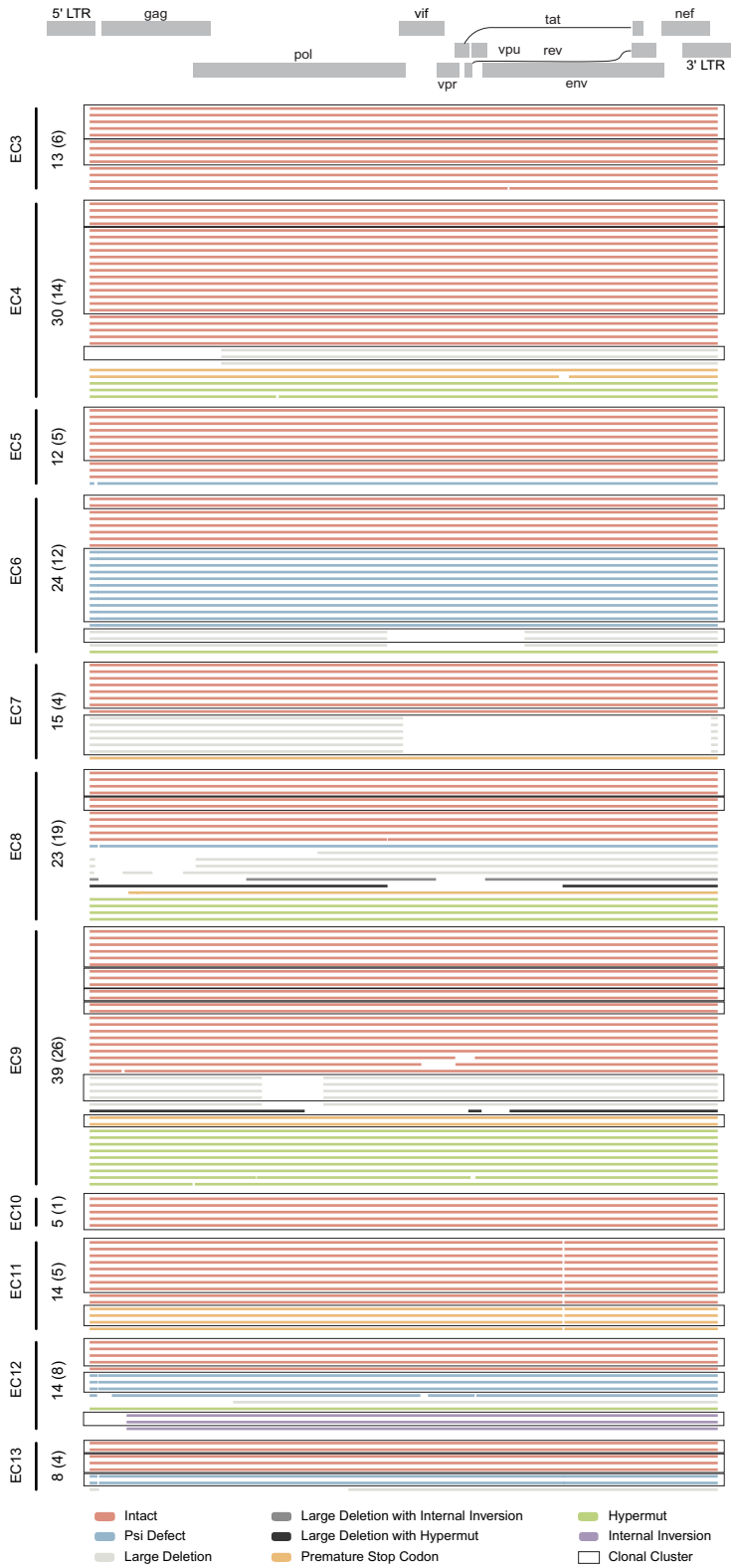
Peer review information *Nature* thanks Nicolas Chomont, Philippe Lemey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



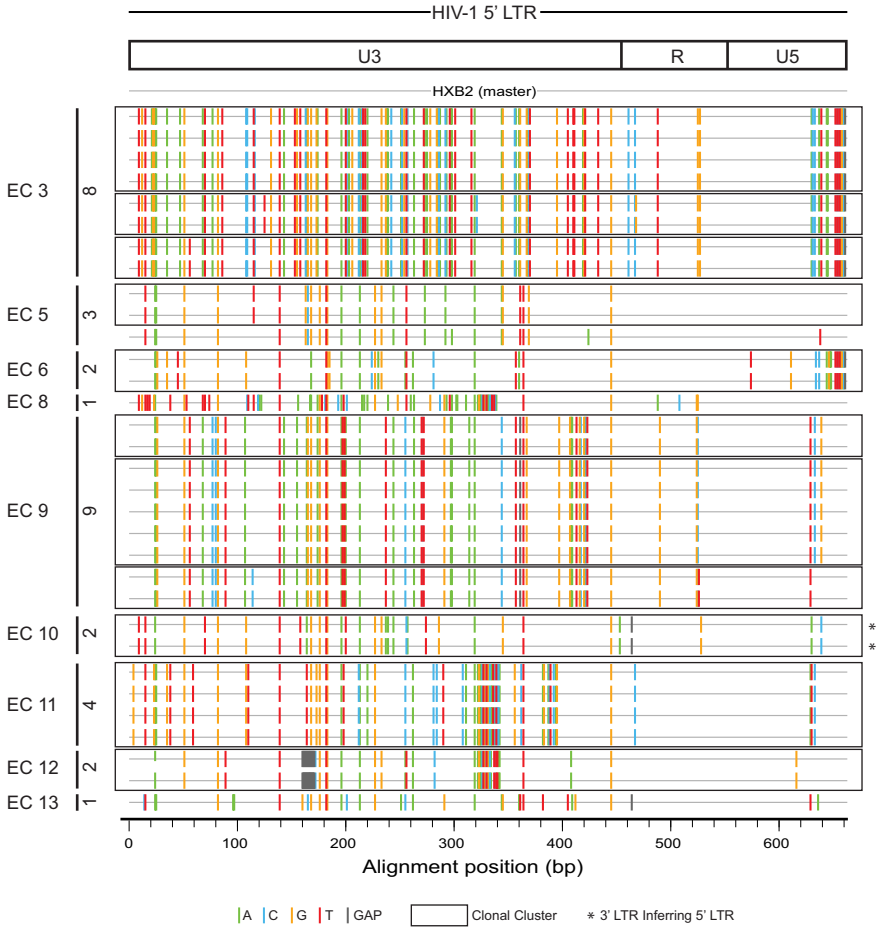
Extended Data Fig. 1 | Viral sequence analysis of intact HIV-1 proviruses from elite controllers. a, Genetic distance (expressed as the average number of base pair substitutions) among all intact near-full-length proviral sequences obtained from each study participant. Clonal sequences were considered to be individual sequences; participants with at least two intact proviruses are included ($n = 175$ intact proviral sequences from 24 elite controllers and $n = 147$ intact proviral sequences from 26 ART-treated individuals). **b**, Frequencies of proviral species (copies per million resting $CD4^+$ T cells) detected by IPDA from EC2. **c**, Proportion of optimal CTL epitopes (restricted by autologous HLA class I isotypes) with wild-type sequences within intact HIV-1 clade B sequences. Each dot represents one intact proviral sequence. $n = 182$ and $n = 133$ HIV-1 clade B intact sequences from 47 elite controllers and 34 ART-treated individuals are included, respectively. Optimal CTL epitopes matching the clade B consensus sequences were considered to be wild-type sequences. Clonal sequences were considered to be individual sequences. **d**, **e**, Average proportions of autologous HLA-class I restricted optimal CTL epitopes with wild-type sequences calculated from intact proviruses in each study participant. Clonal sequences were counted either once (**d**) or as individual sequences (**e**). Each dot represents one study participant. **f**, Proportions of optimal CTL epitopes containing escape variants (restricted by HLA-A01/A02 supertypes, HLA-A03 supertype or HLA-B*27/B*57) within intact proviruses from elite controllers and ART-treated individuals. Each dot represents one intact proviral sequence. Clonal sequences were counted individually. **g**, **h**, Proportion of clonal intact proviruses among all intact proviruses within each study participant (**g**) or within all intact proviruses from elite controllers and ART-treated individuals (**h**). Study participants for whom at least two intact proviruses were detected are included in **g** and **h**. Two-tailed Mann–Whitney *U*-tests were used for data shown in **a**, **c**–**g**; two-sided Fisher's exact test was used for data shown in **h**.



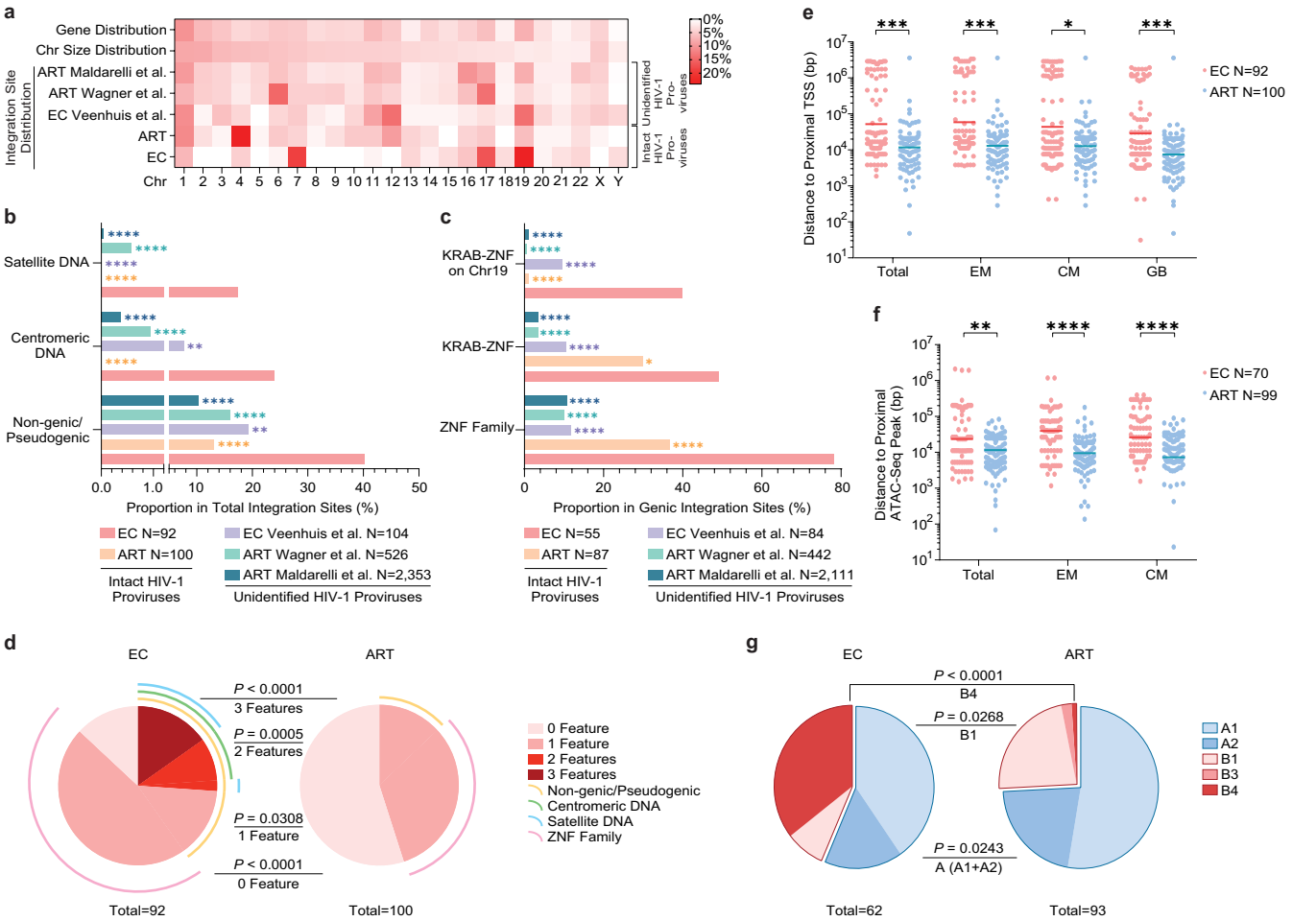
Extended Data Fig. 2 | Longitudinal evolution of CD4⁺ T cell counts and HIV-1 viral loads in EC1–EC13. The recorded diagnosis date of HIV-1 infection for each study participant is shown as the first date on the *x* axis. PBMC sampling time points are indicated by red arrows.



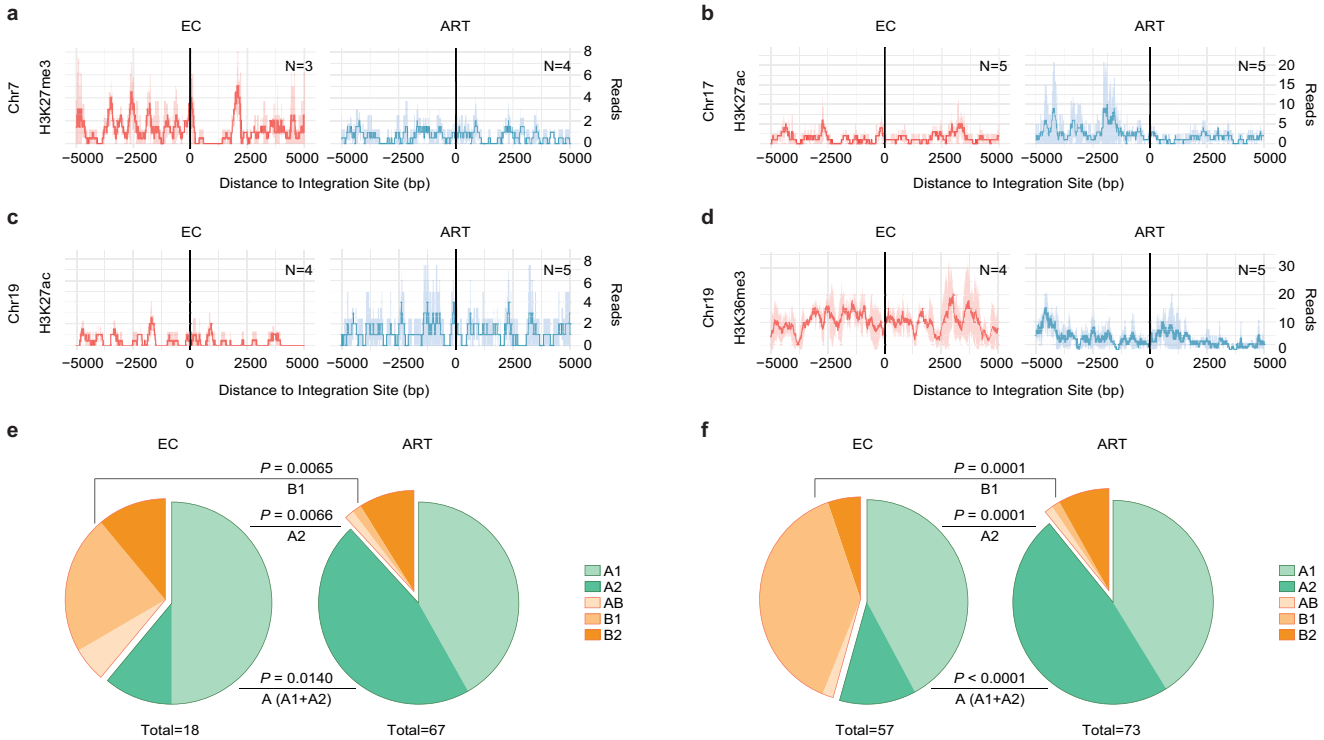
Extended Data Fig. 3 | The structural composition of proviral reservoirs in elite controllers. Virograms reflect the genetic coverage of individual sequences of proviral genomes analysed in EC3–EC13. Numbers of total near-full-length proviral sequences obtained from each individual are shown on the *y* axis; numbers of independent sequences are indicated in brackets. Open boxes indicate clonal clusters.



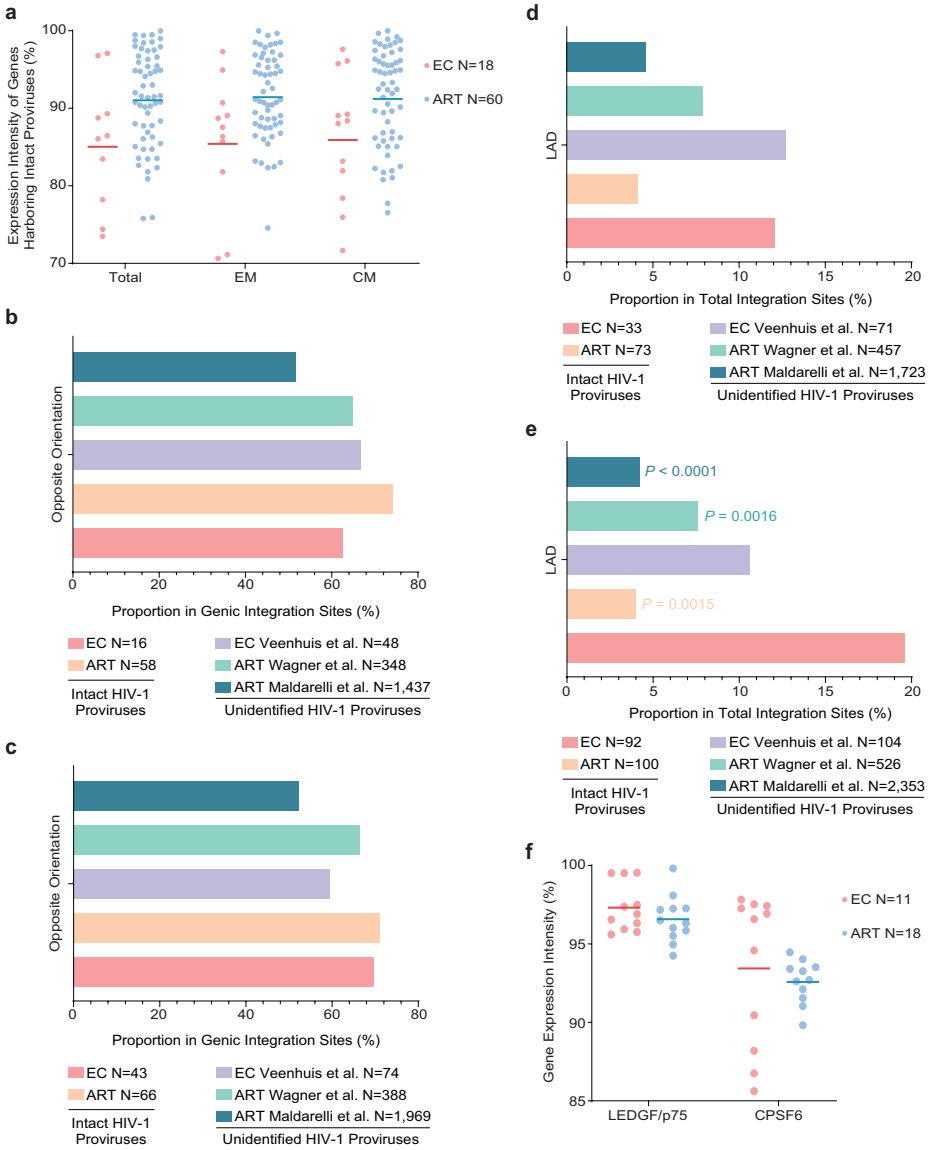
Extended Data Fig. 4 | The variations in HIV-1 DNA sequences in 5′ LTR regions from intact proviruses isolated from the indicated elite controllers, relative to HXB2. Numbers of 5′ LTR sequences of intact proviruses obtained from each individual are shown on the vertical axis. Open boxes indicate clonal clusters.



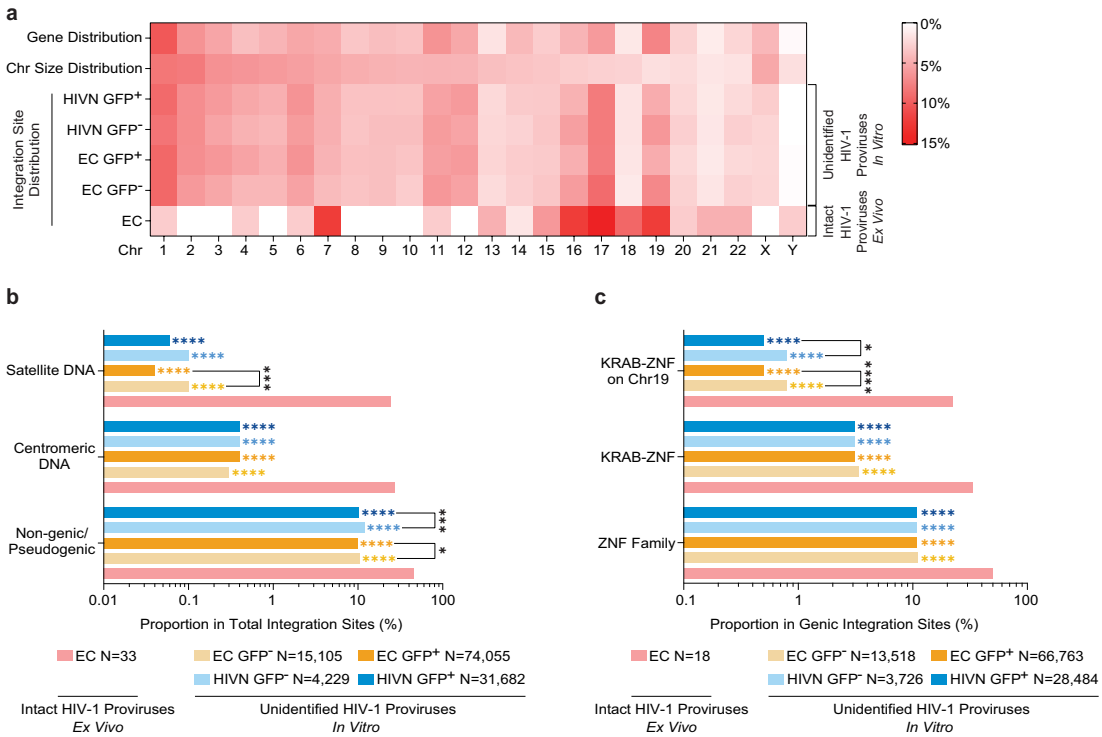
Extended Data Fig. 5 | Features of the chromosomal integration sites of intact proviruses from elite controllers after counting clonal sequences individually. a, Heat map indicating the relative proportion of proviral integration sites of intact proviruses in each chromosome in elite controllers, relative to corresponding data from long-term ART-treated individuals[14]. Proviral integration site data from previous publications[9,15,17] are shown for comparison; integration sites from intact and defective proviruses were not distinguished in these studies. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of near-full-length intact proviruses located in the indicated genomic regions. Data from near-full-length intact proviral sequences in long-term ART-treated individuals are shown as a reference[14]; chromosomal integration sites from unselected (intact and defective) proviral sequences in elite controllers[9] and in ART-treated individuals[15,17] are also shown for comparison. **d**, SPICE diagrams[59] showing the proportion of intact proviruses with the indicated chromosomal integration site features in elite controllers and ART-treated individuals. **e**, **f**, Chromosomal distance between integration sites of intact proviruses and the most proximal transcriptional start sites (determined by RNA-seq) (**e**) or to the most proximal ATAC-seq peak (**f**) in autologous total, central memory and effector memory $CD4^+$ T cells and in the Genome Browser (GB). Horizontal lines show the geometric mean. **g**, Proportions of proviral sequences located in structural compartments A and B, as determined using previously published Hi-C-seq data[29]. Chromosomal integration regions not covered in the previous study[29] were excluded from the analysis. **f**, **g**, Sequences in genomic regions included in the blacklist for functional genomics analysis identified by the ENCODE and modENCODE consortia[28] were excluded owing to the absence of reliable ATAC-seq and Hi-C-seq reads in such repetitive regions. **a**–**g**, All members of clonal clusters were included as individual sequences. $^{****}P < 0.0001$, $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$; FDR-adjusted two-sided Fisher's exact tests were used for data shown in **b** and **c**; two-sided Fisher's exact tests were used for data shown in **d** and **g**; FDR-adjusted two-tailed Mann–Whitney *U*-tests were used for data shown in **e** and **f**; all comparisons were made between elite controllers and reference groups.



Extended Data Fig. 6 | Epigenetic features of the chromosomal integration sites of intact proviruses from elite controllers. a–**d**, Numbers of DNA-sequencing reads associated with activating (H3K27ac) or repressive (H3K27me3) histone protein modifications in proximity to integration sites from elite controllers and long-term ART-treated individuals; median and confidence intervals (defined by one standard deviation) of ChIP–seq data from primary memory CD4$^+$ T cells included in the ROADMAP repository[26] are shown. Negative distances indicate genomic regions upstream of the HIV-1 5′ LTR host–viral junction; positive distances indicate regions downstream of the 3′ LTR viral–host junction. DNA-sequencing reads associated with H3K36me3, a chromatin mark that is atypically enriched in *KRAB-ZNF* genes on chromosome 19, are also shown[29]. **e**, **f**, Proportions of intact proviral sequences located in structural compartments A and B (and associated sub-compartments) by counting clonal sequences once (**e**) or by counting clonal sequences individually (**f**), as determined based on the alignment of chromosomal integration sites of intact proviruses to Hi-C-seq data from Jurkat cells[30]. Chromosomal integration regions not covered in the Jurkat cell study[30] were excluded from the analysis. Compartment B4 was not assessed in the source data[30] for this analysis. Two-sided Fisher's exact tests were used for statistical comparisons; nominal *P* values are reported. **a**–**f**, Sequences in genomic regions included in the blacklist for functional genomics analysis identified by the ENCODE and modENCODE consortia[28] were excluded owing to the absence of reliable ChIP–seq and Hi-C-seq reads in such repetitive regions.



Extended Data Fig. 7 | Accessory features of chromosomal integration sites of intact proviral sequences from elite controllers. a, Expression of host genes that contain intact proviral sequences in elite controllers and long-term ART-treated individuals, as determined by autologous RNA-seq data in total, central memory and effector memory CD4$^+$ T cells. Gene expression percentiles are indicated. **b**, **c**, Orientation of intact proviruses relative to host genes in elite controllers and long-term ART-treated individuals. All data for genic integration sites are included, except for integration sites in genic regions associated with multiple genes in opposing orientations. Integration site data from previous studies of elite controllers[9] and ART-treated individuals[15,17] are shown for comparative purposes. **d**, **e**, Proportion of intact proviruses from elite controllers and long-term ART-treated individuals in lamina-associated domains, determined using Lamin B1–DNA adenine methyltransferase identification (DamID)[61] for resting Jurkat cells. Integration site data from previous studies of elite controllers[9] and ART-treated individuals[15,17] are shown for comparative purposes. **b**, **d**, Clonal proviruses were counted once. **c**, **e**, Clonal proviruses were counted as individual sequences (FDR-adjusted two-sided Fisher's exact tests). **f**, Expression of *LEDGF* (also known as *PSIP1* or *p75*) and *CPSF6* mRNA in autologous total CD4$^+$ T cells from elite controllers and long-term ART-treated individuals, as determined by RNA-seq. Gene expression percentiles are indicated. **a**, **f**, Horizontal lines show the geometric mean. All comparisons were made between elite controllers and reference groups.



Extended Data Fig. 8 | Features of chromosomal integration sites of in vitro-infected CD4$^+$ T cells from elite controllers and HIV-1-negative study participants. a, Heat map showing the relative proportion of proviral integration sites in sorted GFP$^+$ or GFP$^-$ in vitro-infected CD4$^+$ T cells (determined by ligation-mediated PCR[49]) from elite controllers and HIV-1-negative study participants (HIVNs), relative to proviral integration sites of intact proviruses in each chromosome in elite controllers; integration sites from intact and defective proviruses were not distinguished in in vitro-infection studies. Data from GFP$^+$ (n = 74,055) and GFP$^-$ (n = 15,105) CD4$^+$ T cell populations from elite controllers and from GFP$^+$ (n = 31,682) and GFP$^-$ (n = 4,229) CD4$^+$ T cell populations from HIV-1-negative study participants were included. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of proviral integration sites located in indicated genomic regions (**b**) or defined genes (**c**). Data from near-full-length intact proviral sequences in elite controllers are indicated for reference. ****P < 0.0001, ***P < 0.001, *P < 0.05; FDR-adjusted two-sided Fisher's exact tests or two-tailed χ^2 tests were used as appropriate; P values indicating comparisons made between intact proviruses from elite controllers (determined ex vivo) and each in vitro-infection group are shown in corresponding colours.

Extended Data Table 1 | Demographical and clinical characteristics of all study participants

	Elite Controllers (EC)	ART-treated Participants (ART)
Number of participants	64	41
Age in years*	57 (31 - 75)	55 (34 - 73)
Female (%)	18.75%	21.95%
CD4 counts*	908† (450 - 2,282)	726 (316 - 1,649)
Viral loads	Under limit of detection	Under limit of detection
Number of viral load tests*	18 (3 - 91)	32.5 (4 - 73)
HLA-B*27/B*57 (%)	27.34%‡	8.75%
Time since diagnosis (year)*	17 (1 - 34)	17 (5 - 35)
Recorded duration of undetectable viremia (year)*	9 (1 - 24)	9 (2 - 19)

*Median with range.

†$P = 0.0006$, tested using a two-tailed Mann–Whitney *U*-test.

‡$P = 0.0012$, tested using two-sided Fisher's exact test.

nature research

Corresponding author(s): Xu G. Yu

Last updated by author(s): Jun 18, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Quantify One (version 4.4.1), ChemiDoc MP Image Lab software (BioRad, version 6.0.1), BD FACSDiva software (version 8.0.1), QuantaSoft (version 1.7.4.0917)
Data analysis	Los Alamos HIV Sequence Database Hypermut 2.0 (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html), MEGA (https://www.megasoftware.net, version 7.0.26), MUSCLE (http://www.drive5.com/muscle, version 3.8.1551), Graphpad prism (https://www.graphpad.com/scientific-software/prism/, version 8.2.1), UltraCycler v1.0 (Brian Seed and Huajun Wang from MGH CCIB DNACore, unpublished), R (https://www.r-project.org, version 3.5.3), UCSC Genome Browser (https://genome.ucsc.edu), GENCODE (https://www.gencodegenes.org, version 29), Ensembl (https://ensembl.org, version 86), RepeatMasker (www.repeatmasker.org), RSEM (https://deweylab.github.io/RSEM/, version 1.2.22), STAR (https://github.com/alexdobin/STAR,version 2.5.1b), FastQC (https://www.bioinformatics.babraham.ac.uk, version 0.11.9), Trimmomatic (http://www.usadellab.org, version 0.39), Samtools (http://www.htslib.org/, version 1.3.1), MACS2 (https://pypi.python.org/pypi/MACS2, version 2.1.1.20160309), iMethyl (http://imethyl.iwate-megabank.org), ROADMAP (http://www.roadmapepigenomics.org/), MAFFT (https://mafft.cbrc.jp/alignment/software, version 7), Highlighter (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html), SPICE (https://niaid.github.io/spice/), Recombinant Identification Program (https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html), FlowJo software (version 10.6), Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, version 2.2.9), in-house intactness pipeline (https://github.com/BWH-Lichterfeld-Lab/Intactness-Pipeline)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-Seq and ATAC-Seq data have been deposited in a public repository (NCBI GEO, accession number GSE144334). Due to study participant confidentiality concerns, full-length viral sequencing data cannot be publicly released, but will be made available to investigators upon reasonable request and after signing a coded tissue agreement. The Los Alamos HIV Sequence Database Hypermut 2.0 and the Los Alamos HIV Immunology Database 2.0 are available at www.hiv.lanl.gov. The iMethyl database is available at http://imethyl.iwate-megabank.org. ROADMAP epigenomic data are available at http://www.roadmapepigenomics.org.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of n=64 EC and n=41 ART-treated individuals were analyzed in data described in Figure 1. In Figure 2-4, n=11 EC are described in detail. No computational approach was used to determine these sample sizes, testing was based on availability of more than 50 million PBMC per study participant.
Data exclusions	No data from the described individuals were excluded.
Replication	Viral and integration site sequencing was performed once for each individual proviral sequence. To test the accuracy of our sequencing approach, we repeated sequencing of near full-length HIV-1 DNA from the 8E5 cell line 50 consecutive times, which resulted in 100% identical sequences in all runs.
Randomization	No randomization was performed, because we performed a cross-sectional analysis of study participants enrolled in an observational study.
Blinding	Coded samples from study participants were used throughout the study; laboratory personnel was not blinded with regard to the respective study cohorts, since this was a non-interventional, observational study. All sequencing reactions were performed at a local core facilities; core facility employees were fully blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	CD3 (clone OKT3, BioLegend, catalog 317332) CD4 (clone RPA-T4, BioLegend, catalog 300518) CCR7 (clone G043H7, BioLegend, catalog 353216) CCD45RO (clone UCHL1, BioLegend, catalog 304236) CD3/CD8 bi-specific antibody (NIH AIDS Reagent Program #12277)

Validation	CD3 (clone OKT3, BioLegend, catalog 317332): Reactivity: Human Host Species: Mouse Application: FC - Quality tested Application Notes: The OKT3 monoclonal antibody reacts with an epitope on the epsilon-subunit within the human CD3 complex. Clone OKT3 can block the binding of clones SK7 and UCHT1.4 The OKT3 antibody is able to induce T cell activation. Additional reported applications (for the relevant formats) include: immunohistochemical staining of acetone-fixed frozen sections and activation of T cells. The LEAF™ purified antibody (Endotoxin <0.1 EU/µg, Azide-Free, 0.2 µm filtered) is recommended for functional assays (Cat. No. 317304). For highly sensitive assays, we recommend Ultra-LEAF™ purified antibody (Cat. No. 317326) with a lower endotoxin limit than standard LEAF™ purified antibodies (Endotoxin <0.01 EU/µg). Application References: Schlossman S, et al. Eds. 1995. Leucocyte Typing V. Oxford University Press. New York. Knapp W. 1989. Leucocyte Typing IV. Oxford University Press New York. Barclay N, et al. 1997. The Leucocyte Antigen Facts Book. Academic Press Inc. San Diego. Li B, et al. 2005. Immunology 116:487. CD4 (clone RPA-T4, BioLegend, catalog 300518): Reactivity: Human, Chimpanzee Host Species: Mouse Application: FC - Quality tested Application Notes: The RPA-T4 antibody binds to the D1 domain of CD4 (CDR1 and CDR3 epitopes) and can block HIV gp120 binding and inhibit syncytia formation. Additional reported applications (for the relevant formats) include: immunohistochemistry of acetone-fixed frozen sections3,4,5, and blocking of T cell activation1,2. This clone was tested in-house and does not work on formalin fixed paraffin-embedded (FFPE) tissue. The LEAF™ purified antibody (Endotoxin <0.1 EU/µg, Azide-Free, 0.2 µm filtered) is recommended for functional assays (Cat. No. 300516). Application References: Knapp W, et al. 1989. Leucocyte Typing IV. Oxford University Press. New York. (Activ) Moir S, et al. 1999. J. Virol. 73:7972. (Activ) Deng MC, et al. 1995. Circulation 91:1647. (IHC) Friedman T, et al. 1999. J. Immunol. 162:5256. (IHC) Mack CL, et al. 2004. Pediatr. Res. 56:79. (IHC) Lan RY, et al. 2006. Hepatology 43:729. Zenaro E, et al. 2009. J. Leukoc. Biol. 86:1393. (FC) PubMed Yoshino N, et al. 2000. Exp. Anim. (Tokyo) 49:97. (FC) Stoeckius M, et al. 2017. Nat. Methods. 14:865. (PG) CCR7 (clone G043H7, BioLegend, catalog 353216): Reactivity: Human, African Green, Baboon, Cynomolgus, Rhesus Host Species: Mouse Application: FC - Quality tested CCD45RO (clone UCHL1, BioLegend, catalog 304236): Reactivity: Human, Chimpanzee, Cynomolgus, Common Marmoset Host Species: Mouse Application: FC - Quality tested Application Notes: The UCHL1 antibody is commonly used in combination with antibodies against CD45RA to discern memory and naïve T cells. Additional reported applications (for the relevant formats) include: immunohistochemical staining of acetone-fixed frozen tissue sections5 and formalin-fixed paraffin-embedded tissue sections4, Western blotting2, and immunoprecipitation3. Application References: Knapp W, et al. Eds. 1989. Leucocyte Typing IV. Oxford University Press. New York. (FC) Ishii T, et al. 2001. P. Natl. Acad. Sci. USA 98:12138. (WB) Ponsford M, et al. 2001. Clin. Exp. Immunol. 124:315. (IP) Yamada M, et al. 1996. Stroke 27:1155. (IHC) Sakkas LI, et al. 1998. Clin. Diagn. Lab. Immunol. 5:430. (IHC) CD3/CD8 bi-specific antibody (NIH AIDS Reagent Program #12277) The bi-specific CD3/8 (CD3.8) monoclonal antibody was generated by fusing the anti-CD3 mAb producing hybridoma (12F6) with the anti-CD8 mAb producing hymbridoma (OKT8). The resulting anti-CD3/8 antibody, when added to long term peripheral blood co-cultures results in the potent elimination of CD8+ T cells. The remaining cells are highly activated and serve as a reliable source of purified activated cells of interest.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Please see Extended Data Table 1.
Recruitment	EC and ART-treated individuals were recruited based on referral by HIV clinicians and infectious disease physicians. The enrollment protocols allowed recruited of men and women >18 years old, of any race or ethnicity.
Ethics oversight	The Partners Human Research Committee approved all sample collection at MGH and BWH; the IRB of UCSF supervised sample collection at UCSF.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Respiratory disease in rhesus macaques inoculated with SARS-CoV-2

https://doi.org/10.1038/s41586-020-2324-7

Received: 22 March 2020

Accepted: 1 May 2020

Published online: 12 May 2020

Check for updates

Vincent J. Munster[1], Friederike Feldmann[2], Brandi N. Williamson[1], Neeltje van Doremalen[1], Lizzette Pérez-Pérez[1], Jonathan Schulz[1], Kimberly Meade-White[1], Atsushi Okumura[1], Julie Callison[1], Beniah Brumbaugh[3], Victoria A. Avanzato[1], Rebecca Rosenke[2], Patrick W. Hanley[2], Greg Saturday[2], Dana Scott[2], Elizabeth R. Fischer[3] & Emmie de Wit[1]✉

An outbreak of coronavirus disease 2019 (COVID-19), which is caused by a novel coronavirus (named SARS-CoV-2) and has a case fatality rate of approximately 2%, started in Wuhan (China) in December 2019[1,2]. Following an unprecedented global spread[3], the World Health Organization declared COVID-19 a pandemic on 11 March 2020. Although data on COVID-19 in humans are emerging at a steady pace, some aspects of the pathogenesis of SARS-CoV-2 can be studied in detail only in animal models, in which repeated sampling and tissue collection is possible. Here we show that SARS-CoV-2 causes a respiratory disease in rhesus macaques that lasts between 8 and 16 days. Pulmonary infiltrates, which are a hallmark of COVID-19 in humans, were visible in lung radiographs. We detected high viral loads in swabs from the nose and throat of all of the macaques, as well as in bronchoalveolar lavages; in one macaque, we observed prolonged rectal shedding. Together, the rhesus macaque recapitulates the moderate disease that has been observed in the majority of human cases of COVID-19. The establishment of the rhesus macaque as a model of COVID-19 will increase our understanding of the pathogenesis of this disease, and aid in the development and testing of medical countermeasures.

SARS-CoV-2 infection in humans can be asymptomatic or result in mild-to-fatal COVID-19[4–6]. Patients with COVID-19 who develop pneumonia have presented mainly with fever, fatigue, dyspnoea and a cough[7–9]. Rapidly progressing pneumonia—with bilateral opacities on X-rays or patchy shadows and ground-glass opacities by computed tomography scan—has been observed in patients with COVID-19[2,6,10]. Older patients with comorbidities are at the highest risk of an adverse outcome of COVID-19[5,7]. SARS-CoV-2 has been detected in samples from the upper and lower respiratory tracts of patients with COVID-19, in faeces and in blood, but not in urine[5,11–13].

Non-human-primate models that recapitulate aspects of human disease are essential for our understanding of the pathogenic processes involved in severe respiratory disease and for the development of medical countermeasures such as vaccines and antiviral agents.

Clinical respiratory disease

We inoculated eight adult rhesus macaques with the SARS-CoV-2 isolate nCoV-WA1-2020[14]. On day 1 post-inoculation (dpi), all macaques showed changes in their respiratory pattern and piloerection, as reflected in their clinical scores (Fig. 1a). Other signs of disease we observed included reduced appetite, a hunched posture, pale appearance and dehydration (Extended Data Table 1). Coughing was occasionally heard in the room where macaques were housed, but could not be pinpointed to individual macaques. Signs of disease persisted for more than a week, with all macaques being completely recovered between 9 and 17 dpi (Fig. 1a, Extended Data Table 1). We observed weight loss in all macaques (Fig. 1b); body temperatures spiked at 1 dpi but returned to normal levels thereafter (Fig. 1c). Under anaesthesia, the macaques did not show increased respiration; however, all macaques showed irregular respiration patterns (Fig. 1d). Radiographs showed pulmonary infiltrates in all macaques, starting at 1 dpi with mild pulmonary infiltration primarily in the lower lobes of the lung. By 3 dpi, we noted progression of mild pulmonary infiltration into other lung lobes, although these infiltrates were still primarily located in the caudal lung lobes (Fig. 1e). In one macaque, pulmonary infiltrates were observed from 1 to 12 dpi (Extended Data Fig. 1).

Haematological analysis of blood collected during clinical examinations showed a stress leukogram[15] by 1 dpi in the majority of macaques (Extended Data Fig. 2). Lymphocytes and monocytes began to return baseline levels after 1 dpi. Neutrophils had begun to decrease in all macaques by 3 dpi, and continued to decline through to 5 dpi; in 2 of 4 macaques, this led to neutropenia. We observed decreased haematocrit, red blood cell counts and haemoglobin in all macaques at 1 dpi (Extended Data Fig. 2). In addition, reticulocyte percentages and counts had also decreased by this time point. At 5 dpi, 2 of 4 macaques had a normocytic, normochromic nonregenerative anaemia (consistent with the anaemia of a critical illness); these macaques had not returned

[1]Laboratory of Virology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [2]Rocky Mountain Veterinary Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [3]Research Technologies Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. ✉e-mail: emmie.dewit@nih.gov

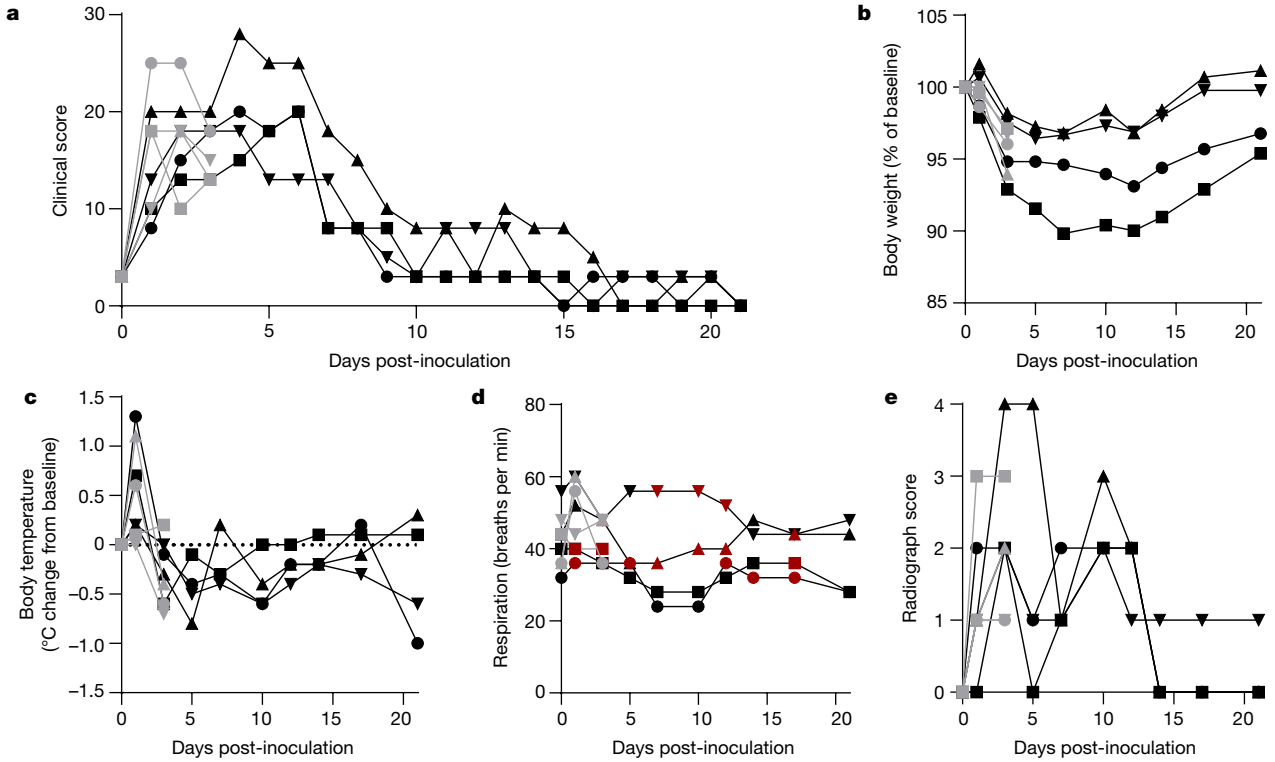


Fig. 1 | Rhesus macaques infected with SARS-CoV-2 develop respiratory disease. a, After inoculation with SARS-CoV-2, macaques were observed for signs of disease, and scored according to a pre-established clinical scoring sheet. **b**, **c**, Body weight (**b**) and body temperature (**c**) were measured in clinical examinations. **d**, Respiration rate was measured, and breathing pattern was recorded; irregular respiration patterns are indicated in red. **e**, Ventrodorsal and lateral radiographs were taken on days on which clinical examination were performed, and were scored for the presence of pulmonary infiltrates: 0, normal; 1, mild interstitial pulmonary infiltrates; 2, moderate pulmonary infiltrates (sometimes with partial effacement of the cardiac border and small areas of pulmonary consolidation); 3, severe interstitial infiltrates, large areas of pulmonary consolidation, alveolar patterns and air bronchograms. Individual lobes were scored, and scores per macaque per day were totalled. Grey, macaques that were euthanized at 3 dpi ($n = 4$); black, macaques that were euthanized at 21 dpi ($n = 4$). The symbols used to denote specific individual macaques are identical throughout the Article.

to their original baseline measurements by 21 dpi. Blood chemistry analysis revealed no values outside of the normal range (Supplementary Table 2).

We analysed serum for changes in cytokine and chemokine levels at different time points after inoculation. We observed statistically significant changes only at 1 dpi (when there were increases in IL-1RA, IL-6, IL-10, IL-15, MCP-1 and MIP-1β) and at 3 dpi, when a small—but statistically significant—decrease in TGFα occurred (Extended Data Fig. 3). Although the levels of some of these cytokines changed at later time points after inoculation, these changes were not statistically significant (Extended Data Fig. 3).

High viral loads in respiratory samples

Virus shedding was highest from the nose (Fig. 2a); virus could be isolated from swabs collected at 1 and 3 dpi, but not thereafter. Viral loads were high in throat swabs immediately after inoculation, but were less consistent than nose swabs thereafter; in one macaque, throat swabs were positive at 1 dpi and at 10 dpi—but not in between (Fig. 2a). One macaque showed prolonged shedding of viral RNA in rectal swabs; infectious virus could not be isolated from these swabs (Fig. 2a) and disease of the intestinal tract (for example, diarrhoea) was not observed. Urogenital swabs remained negative in all macaques throughout the study. For the 4 macaques in the group that was euthanized at 21 dpi, we performed bronchoalveolar lavages at 1, 3 and 5 dpi. We detected high viral loads in fluid from the bronchoalveolar lavage in all macaques at all three time points; infectious virus could be isolated only from bronchoalveolar fluid collected at 1 and 3 dpi (Fig. 2b). No viral RNA was detected in the blood (Fig. 2c) or urine (Fig. 2d).

Interstitial pneumonia

Two groups of 4 macaques were euthanized (one at 3 dpi and the other at 21 dpi), and necropsies were performed. At 3 dpi, varying degrees of lung lesions at the gross pathological scale were observed in all macaques (Fig. 3a, c). At 21 dpi, lesions were visible in the lungs of 2 of 4 macaques (Fig. 3b, c). Additionally, all macaques had an increased ratio of lung weight to body weight (Fig. 3d) as compared to healthy rhesus macaques, indicative of pulmonary oedema. Histologically, 3 of the 4 macaques euthanized at 3 dpi developed some degree of pulmonary pathology. The lesions represented multifocal (Extended Data Fig. 4a), mild-to-moderate interstitial pneumonia that frequently centred on terminal bronchioles. The pneumonia was characterized by thickening of alveolar septa by oedema fluid and fibrin, and small-to-moderate numbers of macrophages and fewer neutrophils. Lungs with moderate changes also had alveolar oedema and fibrin with formation of hyaline membranes. There was minimal type-II pneumocyte hyperplasia. Occasionally, bronchioles showed necrosis, and the loss and attenuation of the epithelium with infiltrates of neutrophils, macrophages and eosinophils. Within the multifocal lesions, there were perivascular infiltrates of small numbers of lymphocytes that formed perivascular cuffs (Extended Data Fig. 4b), and minimal-to-mild, multifocal hyperplasia of bronchiolar-associated lymphoid tissue. Three of 4 macaques at 3 dpi had fibrous adhesions of the lung to the pleura. Histological evaluation showed these adhesions to be composed of mature collagen interspersed with small blood vessels; therefore, this is most probably a chronic change rather than being related to infection with SARS-CoV-2. We observed minimal-to-mild inflammation in the upper airways with multifocal squamous metaplasia of the respiratory epithelium and infiltration of small numbers of neutrophils (Extended Data Fig. 5).

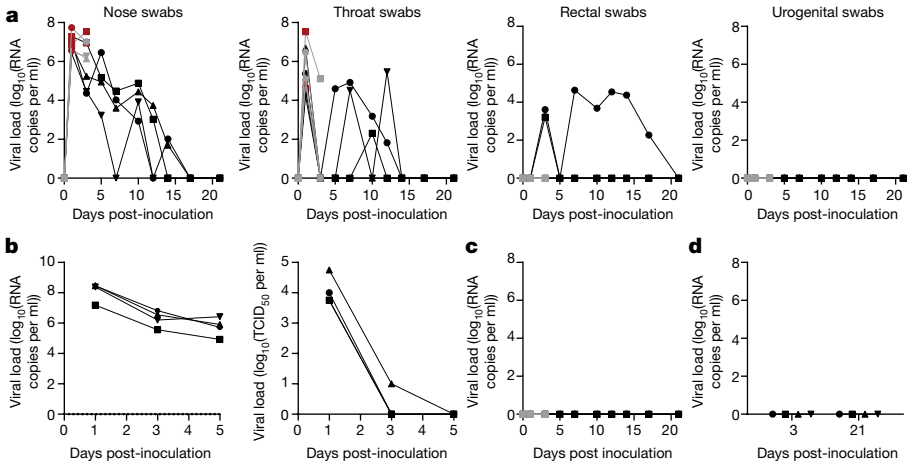


Fig. 2 | Viral loads in respiratory samples and bodily fluids. a, After inoculation, nose, throat, rectal and urogenital swabs were collected; viral loads in these samples were determined by quantitative reverse-transcription PCR. **b**, At 1, 3 and 5 dpi, bronchoalveolar lavages were performed on the 4 macaques that remained in the study through to 21 dpi; viral loads (left) and virus titres (right) were determined in these samples. **c**, **d**, Viral loads were determined in blood collected during clinical examinations (**c**) and urine collected at necropsy at 3 and 21 dpi (**d**). Grey, macaques that were euthanized at 3 dpi ($n = 4$); black, macaques that were euthanized at 21 dpi ($n = 4$); red, virus was isolated from these samples.

Immunohistochemistry using a monoclonal antibody against SARS-CoV demonstrated viral antigen in small numbers of type-I and -II pneumocytes, as well as in alveolar macrophages. We detected antigen-positive macrophages in the mediastinal lymph nodes of three of four macaques (Fig. 3k). We also detected small numbers of antigen-positive lymphocytes and macrophages in the lamina propria of the intestinal tract of all four macaques. In one macaque, all the tissues of the gastrointestinal tract that we collected showed these antigen-positive mononuclear cells (Extended Data Fig. 6).

We performed ultrastructural analysis of lung tissue by transmission electron microscopy, which confirmed the histological diagnosis of interstitial pneumonia. The alveolar interstitial space was greatly expanded by oedema, fibrin, macrophages and neutrophils (Extended Data Fig. 7a). The subepithelial basement membrane was unaffected and maintained a consistent thickness and electron density. We occasionally observed type-I pneumocytes separated from the basement membrane by oedema; the resulting space sometimes contains virions. Affected type-I pneumocytes are lined by small-to-moderate numbers of virions that are 90–160 nm in diameter, with an electron-dense core bound by a less-dense capsid (Extended Data Fig. 7b–e). Alveolar spaces adjacent to affected pneumocytes are filled with a granular, moderately electron-dense material that is consistent with oedema fluid.

Replication in the respiratory tract

All tissues ($n = 37$) collected at necropsy were analysed for the presence of viral RNA. At 3 dpi, high viral loads were detected in the lungs of all macaques (Extended Data Fig. 8a); virus was isolated from the lungs of all four macaques at this time. Additionally, viral RNA was detected in other samples from throughout the respiratory tract (Extended Data Fig. 8), as well as in lymphoid and gastrointestinal tissues. Viral RNA was not detected in other major organs, including the central nervous system. To distinguish viral RNA derived from respiratory secretions from active virus replication, we tested all samples with a presence of viral RNA for the presence of viral mRNA (Extended Data Fig. 8). We detected viral mRNA in all of the respiratory tissues but not in the gastrointestinal tissues (except for the stomach of one macaque), which indicates that virus replication in these tissues is unlikely—although we cannot exclude this possibility, owing to the limited sample size. By 21 dpi, viral RNA—but not mRNA—could still be detected in tissues from all four macaques (Extended Data Fig. 8g).

Serology

We analysed serum for the development of IgG against the SARS-CoV-2 spike protein in an enzyme-linked immunosorbent assay. By 10 dpi, all 4 macaques had seroconverted to the SARS-CoV-2 spike protein; neutralizing responses also started to appear at 10 dpi (Extended Data Fig. 9). The macaque with the lowest and latest neutralizing-antibody response was the one with prolonged viral shedding from the intestinal tract.

Discussion

Clinically, cases of COVID-19 range from being asymptomatic through to mild or severe manifestations[5,6,8,9,13,16]. Patients present with influenza-like symptoms (such as fever and shortness of breath) and may develop pneumonia, requiring mechanical ventilation and support in an intensive care unit[9]. Similar to the diseases caused by infections with SARS-CoV and MERS-CoV, comorbidities such as hypertension and diabetes have an important role in adverse outcomes of COVID-19[8,17,18]. In particular, advanced age and chronic conditions are indicators of a negative outcome[5,7–9,16]—conditions that were absent in our healthy rhesus macaques. An analysis of 1,099 cases of COVID-19 from China has shown that approximately 5% of diagnosed patients developed severe pneumonia that required attending an intensive care unit, 2.3% required mechanical ventilation and 1.4% died[9]. The transient, moderate disease that we observed here in rhesus macaques is thus consistent with the majority of cases of COVID-19 in humans. Pulmonary infiltrates on radiographs—a hallmark of human infection with SARS-CoV-2[2,4,6,7,9,10,16]—were observed in all macaques. The shedding pattern observed in rhesus macaques is notably similar to that previously observed in humans[11,12]. In humans, consistent and high SARS-CoV-2 shedding has been observed from the upper and lower respiratory tract; frequent intermediate shedding has been observed from the intestinal tract; and there has been sporadic detection of SARS-CoV-2 in blood[5]. Similar to humans, the shedding of SARS-CoV-2 in macaques continued after resolution of clinical signs and radiological abnormalities[19]. Limited histopathology is available from patients with COVID-19[20,21]. Our analysis of the histopathological changes observed in the lungs of rhesus macaques suggests that these changes resemble those observed in macaques infected with SARS-CoV and MERS-CoV[21–24], with regard to lesion type and cell tropism.

Serological responses in humans are not typically detectable before 6 days after the onset of symptoms; IgG titres of between

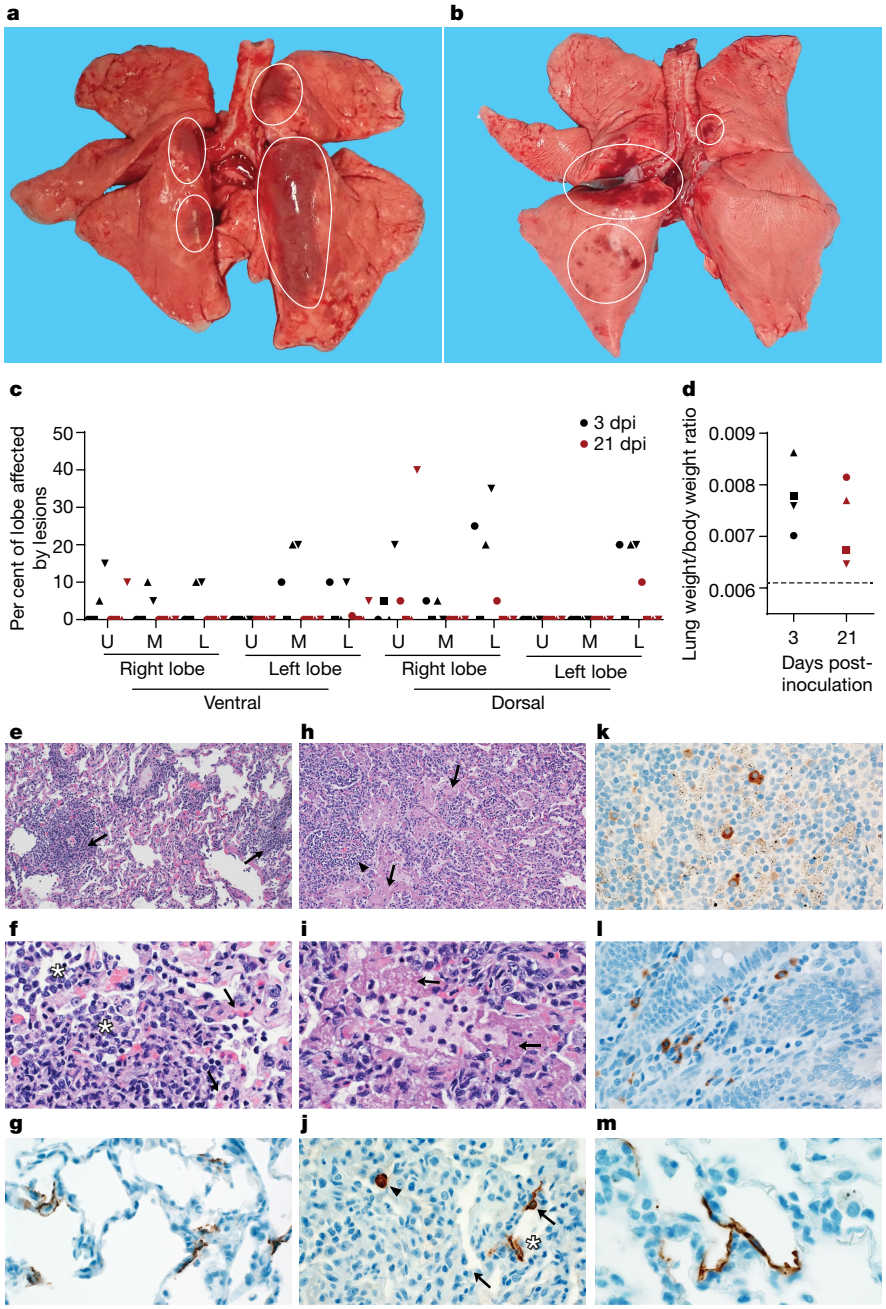


Fig. 3 | Pathological changes in rhesus macaques infected with SARS-CoV-2. Four rhesus macaques were euthanized at 3 and 21 dpi. **a**, **b**, The lungs showed focal areas of hilar consolidation and hyperaemia (circles) at 3 dpi (**a**) and multifocal, random consolidation and hyperaemia (circles) at 21 dpi (**b**). The percentage of the area of the lungs affected by gross lesions was estimated (**c**), and the lung weight-to-bodyweight ratio was calculated (**d**). The dotted line represents the baseline ratio calculated from an in-house collection of lung and body weights from rhesus macaques with normal lungs. **e**–**i**, Histological analysis was performed on tissues collected at 3 dpi. Tissue sections were collected from the same anatomical location for each macaque; three tissue sections were prepared from each of the six lung lobes. In total, 18 lung sections were evaluated for each macaque; representative images are displayed. **e**, Pulmonary vessels surrounded by moderate numbers of lymphocytes, and fewer macrophages (arrows). **f**, Alveoli filled with small-to-moderate numbers of macrophages and neutrophils (asterisks). The adjacent alveolar interstitium (arrows) is thickened by oedema, fibrin, neutrophils, lymphocytes and macrophages. **g**, SARS-CoV-2 antigen detected by immunohistochemistry in type-I pneumocytes. **h**, Pulmonary vessels bounded by lymphocytes (arrowhead) and hyaline membranes (arrows) line the alveolar spaces. **i**, Hyaline membranes line the alveoli (arrows). **j**, SARS-CoV-2 antigen detected by immunohistochemistry in type-I pneumocytes (asterisk) and type-II pneumocytes (arrow), as well as in alveolar macrophages (arrowheads). **k**, SARS-CoV-2 antigen detected by immunohistochemistry in macrophages in a mediastinal lymph node. **l**, SARS-CoV-2 antigen detected by immunohistochemistry in macrophages and lymphocytes in the lamina propria of the caecum. **m**, SARS-CoV-2 detected by immunohistochemistry in type-I pneumocytes. Magnification, 100× (**e**, **h**), 400× (**f**, **g**, **i**–**l**), 1,000× (**m**). U, upper; M, middle; L, lower.

100 and 10,000 have been observed after 12 to 21 days[25,26]. Neutralizing titres were generally between 20 and 160. This corresponds to the results in our rhesus-macaque model, in which IgG responses were detected at around 7–10 dpi. Seroconversion was not directly followed by a decline in viral loads, as observed in patients with COVID-19[5,26].

Together, our rhesus-macaque model recapitulates COVID-19 in humans with regard to virus replication and shedding, the presence of pulmonary infiltrates, histological lesions and seroconversion. This extensive dataset enables us to bridge between the rhesus-macaque model and the disease observed in humans, and to use this model to assess the efficacy of medical countermeasures.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2324-7.

1. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
2. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
3. WHO. *Coronavirus disease (COVID-2019) situation reports*, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/ (2020).
4. Yang, W. et al. Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): a multi-center study in Wenzhou city, Zhejiang, China. *J. Infect.* **80**, 388–393 (2020).
5. Yang, X. et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* **8**, 475–481 (2020).
6. Silverstein, W. K., Stroud, L., Cleghorn, G. E. & Leis, J. A. First imported case of 2019 novel coronavirus in Canada, presenting as mild pneumonia. *Lancet* **395**, 734 (2020).
7. Arentz, M. et al. Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington state. *J. Am. Med. Assoc.* (2020).
8. Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
9. Guan, W. J. et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
10. Shi, H. et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect. Dis.* **20**, 425–434 (2020).
11. Zou, L. et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* **382**, 1177–1179 (2020).
12. Kim, J. Y. et al. Viral load kinetics of SARS-CoV-2 infection in first two patients in Korea. *J. Korean Med. Sci.* **35**, e86 (2020).
13. Tang, A. et al. Detection of novel coronavirus by RT-PCR in stool specimen from asymptomatic child, China. *Emerg. Infect. Dis.* **26**, 1337–1339 (2020).
14. Harcourt, J. et al. Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. *Emerg. Infect. Dis.* **26**, 1266–1273 (2020).
15. Everds, N. E. et al. Interpreting stress responses during routine toxicity studies: a review of the biology, impact, and assessment. *Toxicol. Pathol.* **41**, 560–614 (2013).
16. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *J. Am. Med. Assoc.* (2020).
17. Assiri, A. et al. Epidemiological, demographic, and clinical characteristics of 47 cases of Middle East respiratory syndrome coronavirus disease from Saudi Arabia: a descriptive study. *Lancet Infect. Dis.* **13**, 752–761 (2013).
18. Booth, C. M. et al. Clinical features and short-term outcomes of 144 patients with SARS in the greater Toronto area. *J. Am. Med. Assoc.* **289**, 2801–2809 (2003).
19. Lan, L. et al. Positive RT-PCR test results in patients recovered from COVID-19. *J. Am. Med. Assoc.* (2020).
20. Tian, S. et al. Pulmonary pathology of early-phase 2019 novel coronavirus (COVID-19) pneumonia in two patients with lung cancer. *J. Thorac. Oncol.* **15**, 700–704 (2020).
21. Xu, Z. et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir. Med.* **8**, 420–422 (2020).
22. Ng, D. L. et al. Clinicopathologic, immunohistochemical, and ultrastructural findings of a fatal case of Middle East respiratory syndrome coronavirus infection in the United Arab Emirates, April 2014. *Am. J. Pathol.* **186**, 652–658 (2016).
23. Nicholls, J. M. et al. Lung pathology of fatal severe acute respiratory syndrome. *Lancet* **361**, 1773–1778 (2003).
24. Ding, Y. et al. The clinical pathology of severe acute respiratory syndrome (SARS): a report from China. *J. Pathol.* **200**, 282–289 (2003).
25. Wölfel, R. et al. Virological assessment of hospitalized cases of coronavirus disease 2019. *Nature* **581**, 465–469 (2020).
26. Zhao, Y. et al. Antibody responses to SARS-CoV-2 in patients of novel coronavirus disease 2019. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa344 (2020).

Methods

Because this is an animal model with no prior data, statistical methods could not be used to predetermine sample size.

Ethics and biosafety statement

All experiments in macaques were approved by the Institutional Animal Care and Use Committee of Rocky Mountain Laboratories (National Institutes of Health (NIH)) and carried out by certified staff in an Association for Assessment and Accreditation of Laboratory Animal Care International-accredited facility, according to the institution's guidelines for animal use, following the guidelines and basic principles in the NIH Guide for the Care and Use of Laboratory Animals, the Animal Welfare Act, and the United States Department of Agriculture and the United States Public Health Service Policy on Humane Care and Use of Laboratory Animals. Rhesus macaques were housed in adjacent individual primate cages allowing social interactions, in a climate-controlled room with a fixed light/dark cycle (12-h light/12-h dark). Macaques were monitored at least twice daily throughout the experiment. Commercial monkey chow, treats and fruit were provided twice daily by trained personnel. Water was available ad libitum. Environmental enrichment consisted of a variety of human interaction, manipulanda, commercial toys, videos and music. The Institutional Biosafety Committee (IBC) approved work with infectious SARS-CoV-2 strains under biosafety level 3 conditions. Sample inactivation was performed according to IBC-approved standard operating procedures for removal of specimens from high containment.

Study design

To evaluate the use of rhesus macaques as a model for SARS-CoV-2, eight adult rhesus macaques (4 males, and4 females, age 4–6 years) were inoculated via a combination of intranasal (0.5 ml per nostril), intratracheal (4 ml), oral (1 ml) and ocular (0.25 ml per eye) of a 4×10^5 50% tissue culture infectious dose ($TCID_{50}$) per ml (3×10^8 genome copies per ml) virus dilution in sterile DMEM. The macaques were observed twice daily for clinical signs of disease using a standardized scoring sheet (Supplementary Information, Supplementary Table 1); the same person assessed the macaques throughout the study. The predetermined endpoint for this experiment was 3 dpi for one group of 4 macaques, and 21 dpi for the remaining 4 macaques. Macaques were randomly assigned to a group for necropsy before the start of the experiment. Blinding was not used in this study as all macaques were subjected to the same treatment. Clinical examinations were performed on 0, 1, 3, 5, 7, 10, 12, 14, 17 and 21 dpi on anaesthetized macaques. On exam days, clinical parameters such as bodyweight, body temperature and respiration rate were collected, as well as ventrodorsal and lateral chest radiographs. Chest radiographs were interpreted by a board-certified clinical veterinarian. The following samples were collected at all clinical examinations: nasal, throat, urogenital and rectal swabs, and blood. The total white blood cell count, lymphocyte, neutrophil, platelet, reticulocyte and red blood cell counts, and haemoglobin and haematocrit values were determined from EDTA blood with the IDEXX ProCyte DX analyser (IDEXX Laboratories). Serum biochemistry (albumin, AST, ALT, GGT, BUN and creatinine) was analysed using the Piccolo Xpress Chemistry Analyzer and Piccolo General Chemistry 13 Panel discs (Abaxis). During clinical examinations on 1, 3, and 5 dpi, bronchoalveolar lavages were performed using 10 ml sterile saline. Bronchoalveolar lavages do not induce lung damage when spaced 48 h apart[27,28]. After euthanasia, necropsies were performed. The percentage of gross lung lesions was scored by a board-certified veterinary pathologist and samples of the following tissues were collected: the inguinal lymph node, axillary lymph node, cervical lymph node, salivary gland, conjunctiva, nasal mucosa, oropharynx, tonsil, trachea, all six lung lobes, mediastinal lymph node, right and left bronchus, heart, liver, spleen, pancreas, adrenal gland, kidney, mesenteric lymph node, stomach, duodenum, jejunum, ileum, caecum, colon, urinary bladder, reproductive tract (testes or ovaries depending on sex), bone marrow, frontal brain, cerebellum and brainstem. Histopathological analysis of tissue slides was performed by a board-certified veterinary pathologist blinded to the group assignment of the macaques.

Virus and cells

SARS-CoV-2 isolate nCoV-WA1-2020 (MN985325.1)[14] (Vero passage 3) was provided by the Centers for Disease Control and Prevention, and propagated once in VeroE6 cells in DMEM (Sigma) supplemented with 2% fetal bovine serum (Gibco), 1 mM L-glutamine (Gibco), 50 U/ml penicillin and 50 µg/ml streptomycin (Gibco) (virus isolation medium). The virus stock used was 100% identical to the initial deposited GenBank sequence (MN985325.1) and no contaminants were detected. VeroE6 cells were maintained in DMEM supplemented with 10% fetal calf serum, 1 mM L-glutamine, 50 U/ml penicillin and 50 µg/ml streptomycin. VeroE6 cells were provided by R. Baric and were not authenticated in-house; mycoplasma testing is performed at regular intervals and no mycoplasma has been detected.

Quantitative PCR

RNA was extracted from swabs and bronchoalveolar lavage using the QiaAmp Viral RNA kit (Qiagen) according to the manufacturer's instructions. Tissues (30 mg) were homogenized in RLT buffer and RNA was extracted using the RNeasy kit (Qiagen) according to the manufacturer's instructions. For detection of viral RNA, 5 µl RNA was used in a one-step real-time RT–PCR E assay[29] using the Rotor-Gene probe kit (Qiagen) according to instructions of the manufacturer. In each run, standard dilutions of counted RNA standards were run in parallel, to calculate copy numbers in the samples. For detection of SARS-CoV-2 mRNA, primers targeting open reading frame 7 (ORF7) were designed as follows: forward primer 5′-TCCCAGGTAACAAACCAACC-3′, reverse primer 5′-GCTCACAAGTAGCGAGTGTTAT-3′, and probe FAM-ZEN-CTTGTAGATCTGTTCTCTAAACGAAC-IBFQ. Five µl RNA was used in a one-step real-time RT–PCR using the Rotor-Gene probe kit (Qiagen) according to instructions of the manufacturer. In each run, standard dilutions of counted RNA standards were run in parallel, to calculate copy numbers in the samples.

Histopathology and immunohistochemistry

Histopathology and immunohistochemistry were performed on rhesus macaque tissues. After fixation for a minimum of 7 days in 10% neutral-buffered formalin and embedding in paraffin, tissue sections were stained with haematoxylin and eosin. To detect SARS-CoV-2 antigen, immunohistochemistry was performed using an anti-SARS nucleocapsid protein antibody (Novus Biologicals) at a 1:250 dilution. This antibody was first tested on SARS-CoV-2-infected and uninfected Vero E6 cell pellets, showing specific staining with infected cells and no staining with uninfected cells. The antibody showed specific staining with infected experimental tissue and no staining with uninfected tissue from rhesus macaques. Infected tissue and cell pellet specimens showed no staining when run with rabbit IgG controls (non-specific rabbit IgG substituted for primary antibody). Stained slides were analysed by a board-certified veterinary pathologist.

Transmission electron microscopy. After fixation for 7 days with Karnovsky's fixative at 4 °C, excised tissues were post-fixed for 1 h with 0.5% osmium tetroxide and 0.8% potassium ferricyanide in 0.1 M sodium cacodylate, washed 3 × 5 min with 0.1 M sodium cacodylate buffer, stained 1 h with 1% tannic acid, washed with buffer and then further stained with 2% osmium tetroxide in 0.1 M sodium cacodylate and overnight with 1% uranyl acetate at 4 °C. Specimens were dehydrated with a graded ethanol series with two final exchanges in 100% propylene oxide before infiltration and final embedding in Embed-812 and Araldite resin. Thin sections were cut with a Leica EM UC6 ultramicrotome (Leica),

before viewing at 120 kV on a Tecnai BT Spirit transmission electron microscope (Thermo Fisher/ FEI). Digital images were acquired with a Gatan Rio bottom mount digital camera system (Gatan) and processed using Adobe Photoshop v.CC 2019 (Adobe Systems).

Serum cytokine and chemokine analysis. Serum samples for analysis of cytokine and chemokine levels were inactivated with gamma radiation (2 MRad) according to standard operating procedures. Concentrations of granulocyte colony-stimulating factor, granulocyte–macrophage colony-stimulating factor, IFNγ, IL-1β, IL-1RA, IL-2, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12/23 (p40), IL-13, IL-15, IL-17, MCP-1, MIP-1α, MIP-1β, soluble CD40-ligand (sCD40L), TGFα, TNF, VEGF and IL-18 were measured on a Bio-Plex 200 instrument (Bio-Rad) using the non-human primate cytokine MILLIPLEX map 23-plex kit (Millipore) according to the manufacturer's instructions.

Serology

Sera were analysed by SARS-CoV-2 spike (S) protein enzyme-linked immunosorbent assay (ELISA), as done previously for MERS-CoV[30]. In brief, maxisorp (Nunc) plates were coated overnight with 100 ng/well S protein diluted in PBS[31] (a gift of B. Graham) and blocked with blocker casein in PBS (Life Technologies). Sera were serially diluted in duplicate. SARS-CoV-2-specific antibodies were detected using anti-monkey IgG polyclonal antibody HRP-conjugated antibody (KPL), peroxidase-substrate reagent (KPL) and stop reagent (KPL). Optical density (OD) was measured at 405 nm. The threshold of positivity was calculated by taking the average of the day-0 values multiplied by 3.

For neutralization, sera were heat-inactivated (30 min, 56 °C) and twofold serial dilutions were prepared in 2% DMEM. Then, 100 $TCID_{50}$ of SARS-CoV-2 was added. After 60 min incubation at 37 °C, virus:serum mixture was added to VeroE6 cells and incubated at 37 °C and 5% CO_2. At 5 dpi, cytopathic effect was scored. The virus neutralization titre is expressed as the reciprocal value of the highest dilution of the serum that still inhibited virus replication. All sera were analysed in duplicate.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data have been deposited in Figshare at https://doi.org/10.35092/yhjc.12026910.

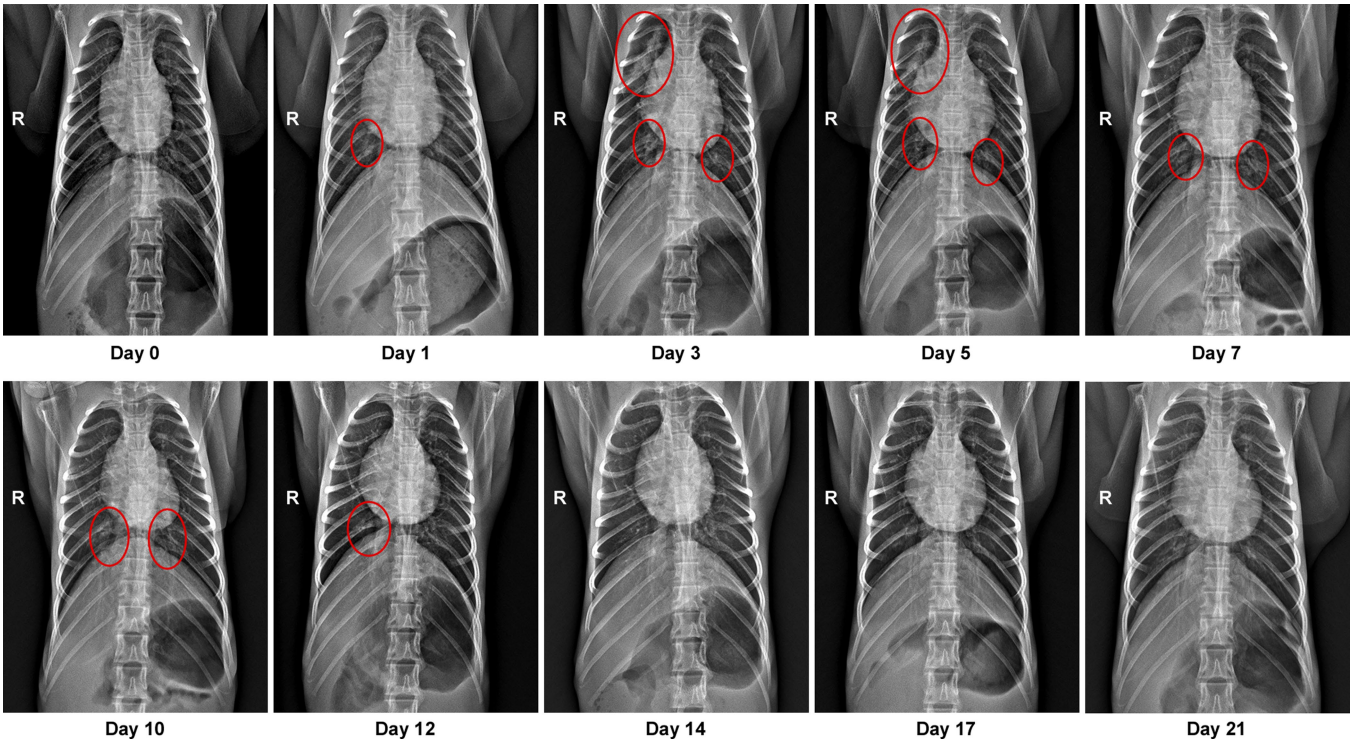
27. Haley, P. J., Muggenburg, B. A., Rebar, A. H., Shopp, G. M. & Bice, D. E. Bronchoalveolar lavage cytology in cynomolgus monkeys and identification of cytologic alterations following sequential saline lavage. *Vet. Pathol.* **26**, 265–273 (1989).
28. Krombach, F. et al. Short-term and long-term effects of serial bronchoalveolar lavages in a nonhuman primate model. *Am. J. Respir. Crit. Care Med.* **150**, 153–158 (1994).
29. Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* **25**, (2020).
30. van Doremalen, N. et al. High prevalence of Middle East respiratory coronavirus in young dromedary camels in Jordan. *Vector Borne Zoonotic Dis.* **17**, 155–159 (2017).
31. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).

Acknowledgements We thank S. Gerber and N. Thornburg for providing the SARS-CoV-2 isolate used in this study; B. Graham, K. Corbett and O. Abiona for providing spike protein for serology; A. Mora for help with figure design. and staff of the Rocky Mountain Veterinary Branch (NIAID (NIH)) for animal care. This study was supported by the Intramural Research Program (NIAID (NIH)).

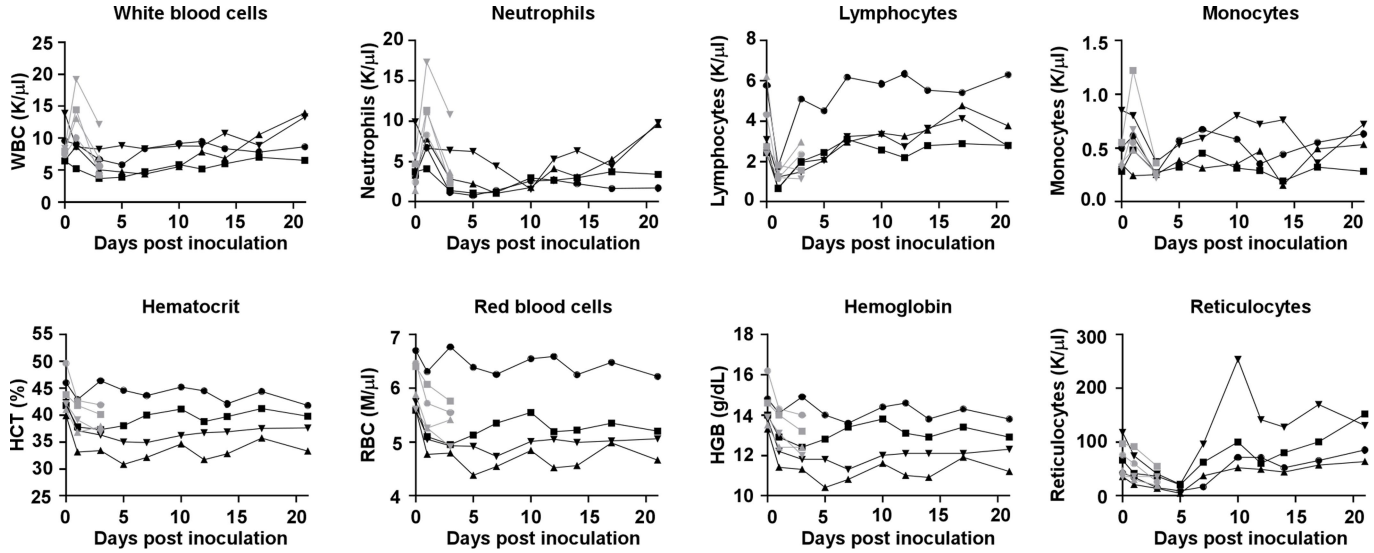
Author contributions V.J.M. and E.d.W. designed the study; V.J.M., F.F., B.N.W., N.v.D., L.P.P., J.S., K.M.-W., A.O., J.C., B.B., V.A.A., R.R., P.W.H., G.S., E.R.F., D.S. and E.d.W. acquired, analysed and interpreted the data; V.J.M., P.W.H., E.R.F., D.S. and E.d.W. wrote the manuscript. All authors have approved the submitted version.

Competing interests The authors declare no competing interests.

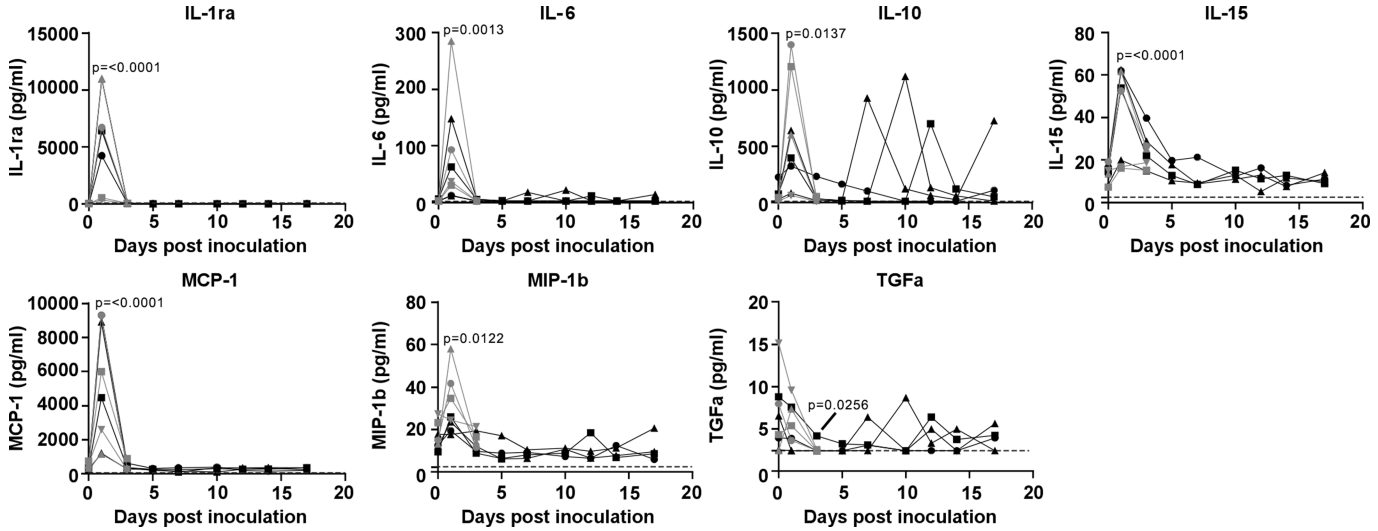
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-020-2324-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2324-7.
Correspondence and requests for materials should be addressed to E.d.W.
Peer review information *Nature* thanks Wolfgang Baumgärtner, Menno D. de Jong and Patricia Pesavento for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



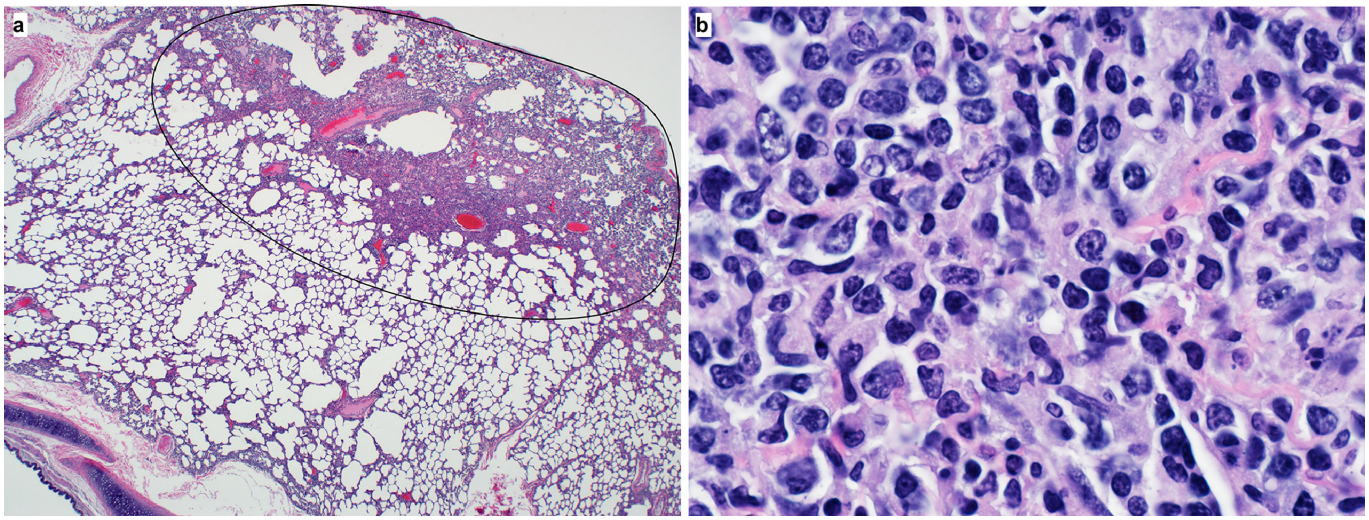
Extended Data Fig. 1 | Pulmonary infiltrates in a rhesus macaque after inoculation. Radiographs show the progression of pulmonary infiltrates throughout the study in a single macaque. This macaque is the one denoted by a black triangle throughout the Article. Circles indicate areas of mild-to-moderate pulmonary infiltrates. R, right side of the macaque. Three chest radiographs were taken at each time point: right lateral, left lateral and ventrodorsal. Only the ventrodorsal radiograph is shown.



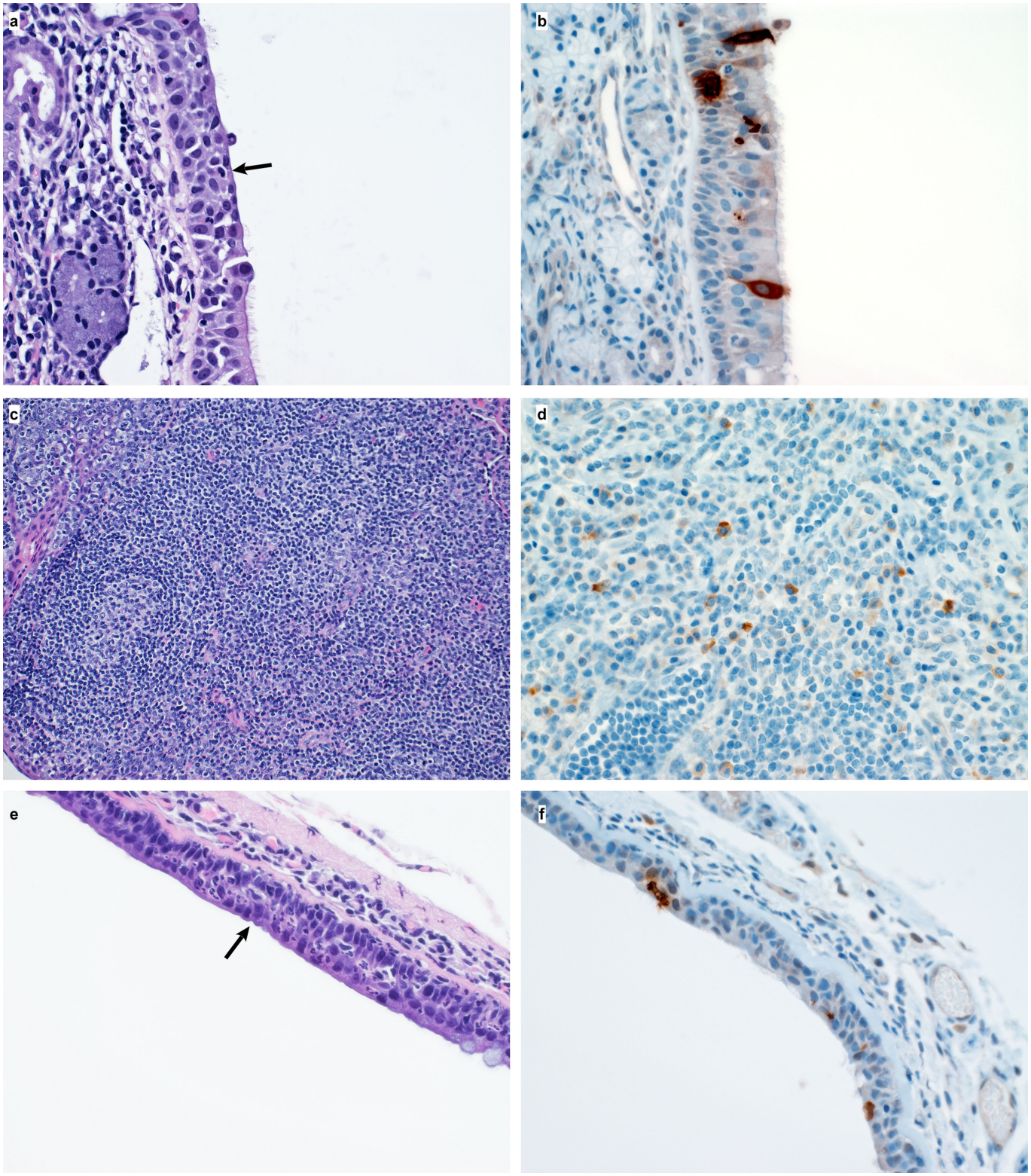
Extended Data Fig. 2 | Haematological changes in rhesus macaques infected with SARS-CoV-2. $n = 8$ macaques at 0, 1 and 3 dpi, and $n = 4$ macaques thereafter.



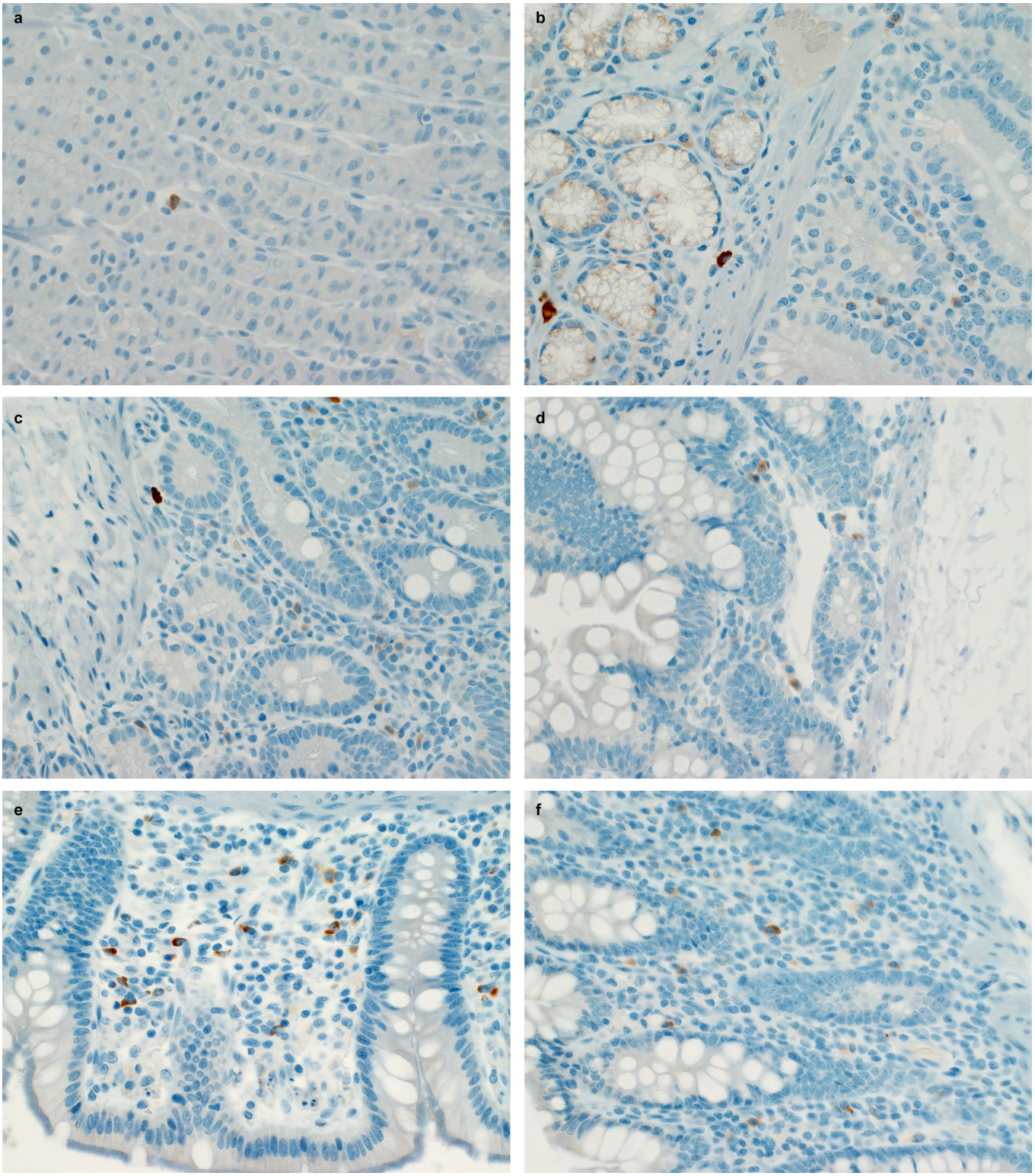
Extended Data Fig. 3 | Cytokine and chemokine levels in the serum of rhesus macaques infected with SARS-CoV-2. The levels of 23 cytokines and chemokines were determined in serum at different time points after inoculation. Levels are displayed only for those cytokines and chemokines for which statistically significant (one-way analysis of variance) differences were observed compared to levels on the day of inoculation. The lower limit of detection is indicated with a dotted line. Serum samples were analysed in duplicate from each macaque for each time point; $n = 8$ macaques at 0, 1, and 3 dpi and $n = 4$ macaques thereafter.



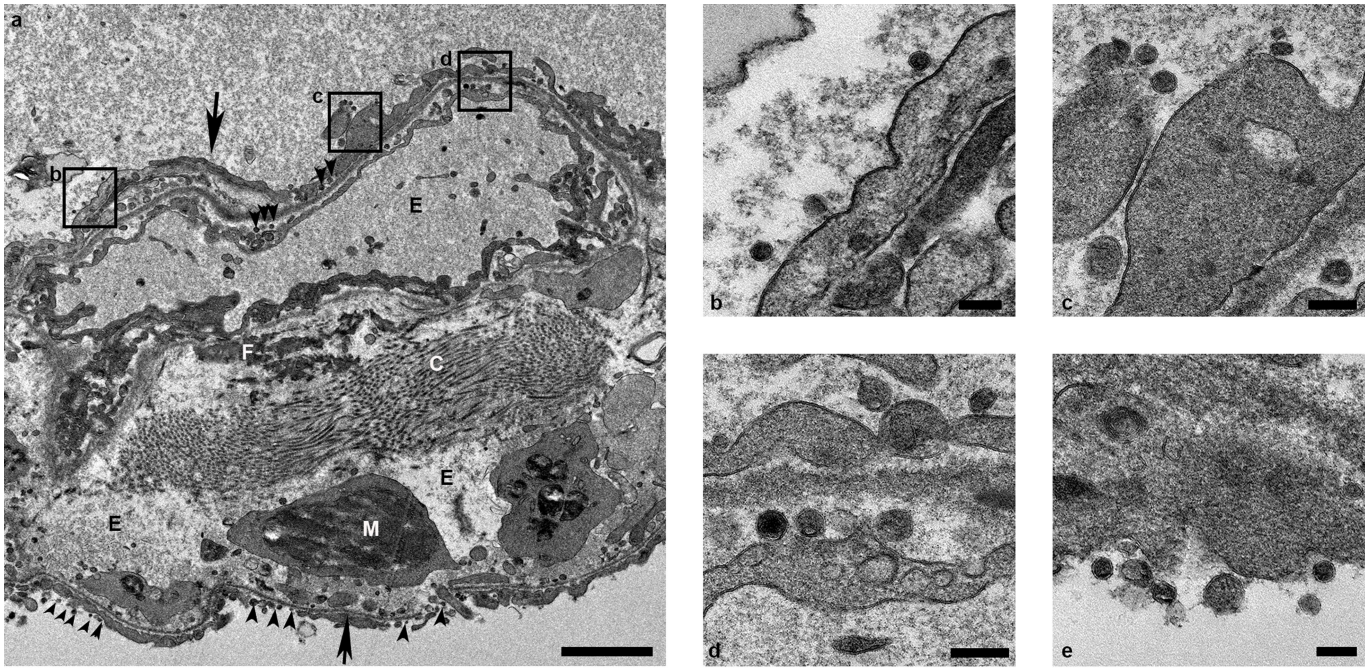
Extended Data Fig. 4 | Histological lesions in the lungs of a rhesus macaque infected with SARS-CoV-2. a, This low-magnification figure displays the focal nature of SARS-CoV-2 lesions in the lungs of macaques euthanized at 3 dpi. The circle indicates the lung affected by lesion; the remaining lung tissue is healthy. **b**, Lymphocytes surround pulmonary vessels. Magnification, 500×. Tissue sections were collected from the same anatomical location for each macaque; three tissue sections were prepared from each of the six lung lobes. In total, 18 lung sections were evaluated for each macaque ($n = 4$); representative images are displayed.



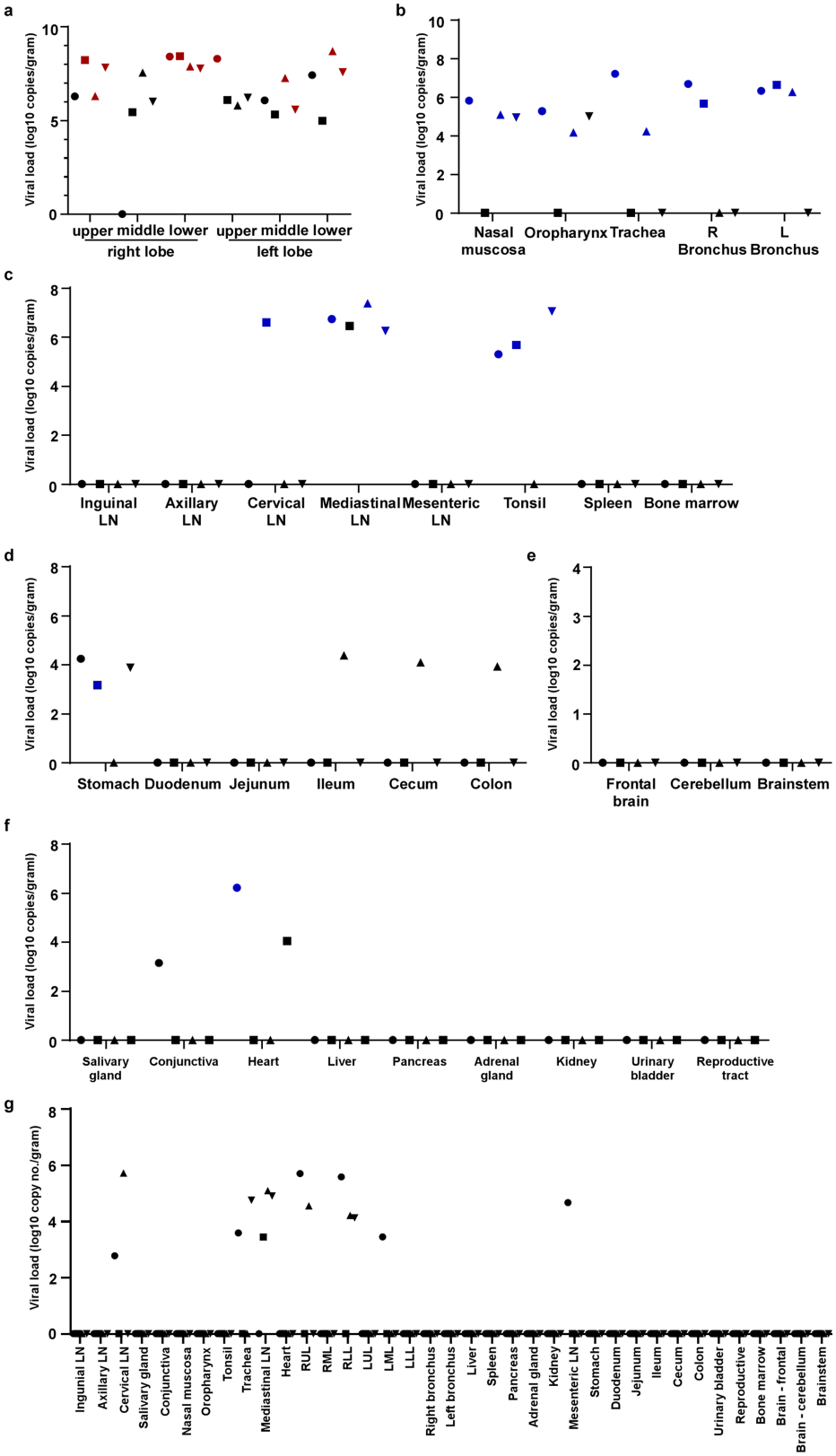
Extended Data Fig. 5 | Histological changes in the respiratory tract of rhesus macaques infected with SARS-CoV-2. a, Squamous metaplasia (arrow) of respiratory epithelium of the nasal turbinate. Magnification, 400×. **b**, SARS-CoV-2 antigen is detected by immunohistochemistry in respiratory epithelium of the nasal turbinate. Magnification, 400×. **c**, Essentially normal tonsil. Magnification, 400×. **d**, SARS-CoV-2 antigen is detected by immunohistochemistry in tonsillar macrophages. Magnification, 400×. **e**, Squamous metaplasia of tracheal columnar epithelium (arrow). Magnification, 400×. **f**, SARS-CoV-2 antigen is detected by immunohistochemistry in tracheal columnar epithelium. Magnification, 400×. Tissue sections were collected from the same anatomical location for each macaque ($n = 4$) and organ; one tissue section was evaluated of the nasal turbinates of each macaque; three tissue sections were evaluated from the tonsil and trachea.



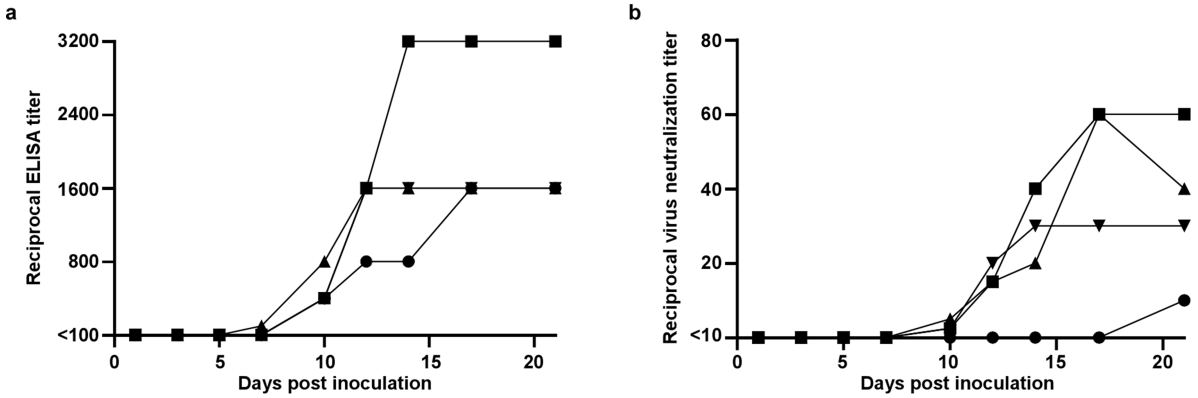
Extended Data Fig. 6 | SARS-CoV-2 antigen in the gastrointestinal tract of a rhesus macaque infected with SARS-CoV-2. a–**f**, Mononuclear cells stained positive for SARS-CoV-2 antigen in the lamina propria of the stomach (**a**), duodenum (**b**), jejunum (**c**), ileum (**d**), caecum (**e**) and colon (**f**) of a macaque infected with SARS-CoV-2 and euthanized on 3 dpi. Tissue sections were collected from the same anatomical location for each macaque ($n = 4$) and organ; three tissue sections were evaluated from each macaque and organ.



Extended Data Fig. 7 | Ultrastructural analysis of the lungs of rhesus macaques infected with SARS-CoV-2. a–**e**, Lung tissue collected on 3 dpi was analysed by transmission electron microscopy. The alveolar interstitium is expanded by oedema (E), fibrin (F) and mononuclear inflammatory cells (M) (**a**). Normal collagen fibres (**c**) and multiple virions (arrowheads) line type-I pneumocytes (arrows). Boxes in **a** indicate areas enlarged in **b**–**d**. **b**–**e**, SARS-CoV-2 virions lining type-I pneumocytes. Scale bars, 2 µm (**a**), 0.2 µm (**b**–**e**). Three tissue samples were collected from each macaque ($n = 4$) and cut into 6 samples for analysis; a minimum of 2 samples were analysed per macaque ($n = 4$).



Extended Data Fig. 8 | Viral loads in tissues collected from rhesus macaques infected with SARS-CoV-2. Eight adult rhesus macaques were inoculated with the SARS-CoV-2 isolate nCoV-WA1-2020 and euthanized at 3 ($n = 4$) or 21 ($n = 4$) dpi. Thirty-seven tissues were collected at necropsy and analysed for the presence of viral RNA by quantitative reverse-transcription PCR (qRT–PCR). **a**–**g**, Tissues are grouped by lung lobes collected at 3 dpi (**a**) (red symbols indicate tissues from which virus could be isolated in Vero E6 cells); other tissues from the respiratory tract at 3 dpi (**b**); lymphoid tissues at 3 dpi (**c**); gastrointestinal tissues at 3 dpi (**d**); the central nervous system at 3 dpi (**e**); the remaining tissues at 3 dpi (**f**); and all tissues collected at 21 dpi (**g**). Blue symbols in **b**–**g** indicate that viral mRNA was also detected in these tissues. L, left; LLL, left lower lung lobe; LML, left middle lung lobe; LN, lymph node; LUL, left upper lung lobe; R, right; RLL, right lower lung lobe; RML, right middle lung lobe; RUL, right upper lung lobe.



Extended Data Fig. 9 | Antibody response in rhesus macaques infected with SARS-CoV-2. a, **b**, Sera collected after inoculation were tested for the presence of IgG against SARS-CoV-2 spike in an ELISA (**a**) and for the presence of neutralizing antibodies in a microneutralization assay (**b**). All sera were analysed in duplicate.

Extended Data Table 1 | Clinical signs observed in rhesus macaques inoculated with SARS-CoV-2

Animal	Clinical signs observed 1-6 dpi	Clinical signs observed 7-21 dpi	Observations at necropsy*
RM1	Hunched posture; piloerection; tachypnea; flushed appearance; red eyes; very agitated; reduced appetite; mildly dehydrated. Euthanized 3 dpi.	N/A	Gross lung lesions. Enlarged tonsils and mediastinal lymph nodes. Fluid-filled stomach, small and large intestine.
RM2	Piloerection; dyspnea; reduced appetite. Euthanized 3 dpi.	N/A	Fluid-filled stomach, small and large intestine.
RM3	Piloerection; tachypnea; flushed appearance; reduced appetite; mildly dehydrated. Euthanized 3 dpi.	N/A	Epistaxis. Gross lung lesions. Enlarged mediastinal lymph nodes. Fluid-filled stomach, small and large intestine.
RM4	Hunched posture; piloerection; tachypnea; dyspnea; reduced appetite. Euthanized 3 dpi.	N/A	Gross lung lesions. Foamy exudate from trachea. Enlarged mediastinal lymph nodes. Fluid-filled stomach, small and large intestine.
RM5	Hunched posture; piloerection; tachypnea; dyspnea; reduced appetite.	Tachypnea; dyspnea; reduced appetite; mildly dehydrated. Recovered on 9 dpi.	Gross lung lesions. Enlarged mesenteric lymph nodes.
RM6	Hunched posture; piloerection; tachypnea; dyspnea; reduced appetite; serous nasal discharge.	Piloerection; bradypnea; mildly dehydrated; crusty nasal discharge. Recovered on 10 dpi.	None.
RM7	Hunched posture; piloerection; pale appearance; tachypnea; dyspnea; irregular; labored respirations; anorexia; mildly dehydrated; serous nasal discharge.	Hunched posture; piloerection; pale appearance; tachypnea; dyspnea; reduced appetite; mildly dehydrated; crusty nasal discharge. Recovered on 17 dpi.	None.
RM8	Hunched posture; piloerection; pale appearance; increased, dyspnea; reduced appetite; serous nasal discharge.	Hunched posture; piloerection; pale appearance; increased, dyspnea; nasal discharge; reduced appetite; mildly dehydrated; serous nasal discharge. Recovered on 13 dpi.	Gross lung lesions.

*Incidental observations not related to infection with SARS-CoV-2 were omitted from this table.

nature research

Corresponding author(s): Emmie de Wit

Last updated by author(s): Mar 28, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Data were analyzed using Graphpad Prism 8.2.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data have been deposited in Figshare: https://doi.org/10.35092/yhjc.12026910

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Since this is a model with no prior data, it was not possible to perform a power analysis. The sample size was based on experience with other nonhuman primate models of respiratory disease.
Data exclusions	No data were excluded.
Replication	Lung histology: for each animal (n=4), 3 sections were evaluated from all 6 lung lobes. Gastrointestinal tract, trachea, tonsil histology: Tissue sections were collected from the same anatomical location for each animal (n=4) and organ; three tissue sections were evaluated from each animal and organ. Nasal turbinate histology: Tissue sections were collected from the same anatomical location for each animal (n=4) and organ; one tissue section was evaluated from each animal and organ. Radiographs: Three chest radiographs were taken from each animal at each clinical exam: right-lateral, left-lateral and ventro-dorsal; only the ventro-dorsal radiograph is shown. Cytokine analysis: serum samples were analyzed in duplicate from each animal for each timepoint; n= 8 animals on 0, 1, and 3 dpi and n=4 animals thereafter. Ultrastructural analysis: Three tissue samples were collected from each animal (n=4) and cut into 6 samples for analysis; a minimum of 2 samples were analyzed per animal (n=4). Serological analysis: Serum samples were analyzed in duplicate from each animal (n=4)for each timepoint.
Randomization	Animals were randomly assigned to the group euthanized at 3 dpi or 21 dpi.
Blinding	Blinding was not used since there was only a single treatment (inoculation with SARS-CoV-2).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-SARS nuclocapsid protein antibody; Novus Biologicals, cat.no. NB100-56576, lotno.111003003 anti-monkey IgG (gamma) antibody, peroxidase-labeled; KPL, cat.no. 5220-0333, lot no. 10329492
Validation	Validation of cross-reactivity of SARS-CoV to SARS-CoV-2 in IHC was done in-house by embedding SARS-CoV-2 infected Vero cells in histogel and producing and staining histology slides.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	VeroE6: Ralph Baric, University of North Carolina, Chapel Hill, USA
Authentication	Not authenticated in-house.
Mycoplasma contamination	Mycoplasma testing confirmed negative at regular intervals.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Rhesus macaques, Chinese origin, adult (4-6 years), 4 males, 4 females
Wild animals	No wild animals were used.
Field-collected samples	No samples were collected in the field.
Ethics oversight	All animal experiments were approved by the Institutional Animal Care and Use Committee of Rocky Mountain Laboratories, NIH and carried out by certified staff in an Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International accredited facility, according to the institution's guidelines for animal use, following the guidelines and basic principles in the NIH Guide for the Care and Use of Laboratory Animals, the Animal Welfare Act, United States Department of Agriculture and the United States Public Health Service Policy on Humane Care and Use of Laboratory Animals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2

https://doi.org/10.1038/s41586-020-2423-5

Received: 23 April 2020

Accepted: 2 June 2020

Published online: 9 June 2020

Check for updates

Brandi N. Williamson[1], Friederike Feldmann[2], Benjamin Schwarz[3], Kimberly Meade-White[1], Danielle P. Porter[4], Jonathan Schulz[1], Neeltje van Doremalen[1], Ian Leighton[3], Claude Kwe Yinda[1], Lizzette Pérez-Pérez[1], Atsushi Okumura[1], Jamie Lovaglio[2], Patrick W. Hanley[2], Greg Saturday[2], Catharine M. Bosio[3], Sarah Anzick[5], Kent Barbian[5], Tomas Cihlar[4], Craig Martens[5], Dana P. Scott[2], Vincent J. Munster[1] & Emmie de Wit[1]✉

Effective therapies to treat coronavirus disease 2019 (COVID-19) are urgently needed. While many investigational, approved, and repurposed drugs have been suggested as potential treatments, preclinical data from animal models can guide the search for effective treatments by ruling out those that lack efficacy in vivo. Remdesivir (GS-5734) is a nucleotide analogue prodrug with broad antiviral activity[1,2] that is currently being investigated in COVID-19 clinical trials and recently received Emergency Use Authorization from the US Food and Drug Administration[3,4]. In animal models, remdesivir was effective against infection with Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV)[2,5,6]. In vitro, remdesivir inhibited replication of SARS-CoV-2[7,8]. Here we investigate the efficacy of remdesivir in a rhesus macaque model of SARS-CoV-2 infection[9]. Unlike vehicle-treated animals, macaques treated with remdesivir did not show signs of respiratory disease; they also showed reduced pulmonary infiltrates on radiographs and reduced virus titres in bronchoalveolar lavages twelve hours after the first dose. Virus shedding from the upper respiratory tract was not reduced by remdesivir treatment. At necropsy, remdesivir-treated animals had lower lung viral loads and reduced lung damage. Thus, treatment with remdesivir initiated early during infection had a clinical benefit in rhesus macaques infected with SARS-CoV-2. Although the rhesus macaque model does not represent the severe disease observed in some patients with COVID-19, our data support the early initiation of remdesivir treatment in patients with COVID-19 to prevent progression to pneumonia.

We have recently established a rhesus macaque model of SARS-CoV-2 infection[9]. In this model, infected rhesus macaques develop mild to moderate, transient respiratory disease, with pulmonary infiltrates visible on radiographs and a shedding pattern similar to that observed in patients with COVID-19. The observed clinical signs and high viral loads enable the testing of the treatment efficacy of direct-acting antivirals in this model.

Distribution of remdesivir to the lungs

Two groups of six rhesus macaques were inoculated with SARS-CoV-2 strain nCoV-WA1-2020. Twelve hours after inoculation, one group was given 10 mg kg^{-1} intravenous remdesivir and the other group was treated with an equal volume of vehicle solution (2 ml kg^{-1}). Treatment was continued 12 h after the first treatment and every 24 h thereafter with a dose of 5 mg kg^{-1} remdesivir or an equal volume of vehicle solution (1 ml kg^{-1}). The concentration of remdesivir was determined in serum collected 12 h after the initial treatment and 24 h after subsequent doses (immediately before the next dose of treatment was administered). Remdesivir (prodrug GS-5734), its downstream alanine metabolite (GS-704277) and the parent nucleoside (GS-441524) were detected in serum from all remdesivir-treated animals (Extended Data Fig. 1a). Serum levels of the prodrug and downstream metabolites were consistent with previously published plasma levels of these compounds in healthy rhesus macaques, which showed a short systemic half-life for GS-5734 (0.39 h) resulting in transient conversion to the intermediate GS-704277 and persistence of the downstream GS-441524 product at higher plasma levels[10].

Concentrations of the metabolite GS-441524 were measured in lung tissue collected from each lung lobe seven days post-inoculation (dpi) and 24 h after the last dose of remdesivir was administered; the metabolite was readily detectable in all remdesivir-treated animals. GS-441524 was generally distributed throughout all six lobes of the lung (Extended Data Fig. 1b). GS-704277 was not detected in lung tissue. Although the pharmacologically active metabolite of remdesivir is

[1]Laboratory of Virology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [2]Rocky Mountain Veterinary Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [3]Laboratory of Bacteriology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [4]Gilead Sciences, Foster City, CA, USA. [5]Research Technologies Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. ✉e-mail: emmie.dewit@nih.gov

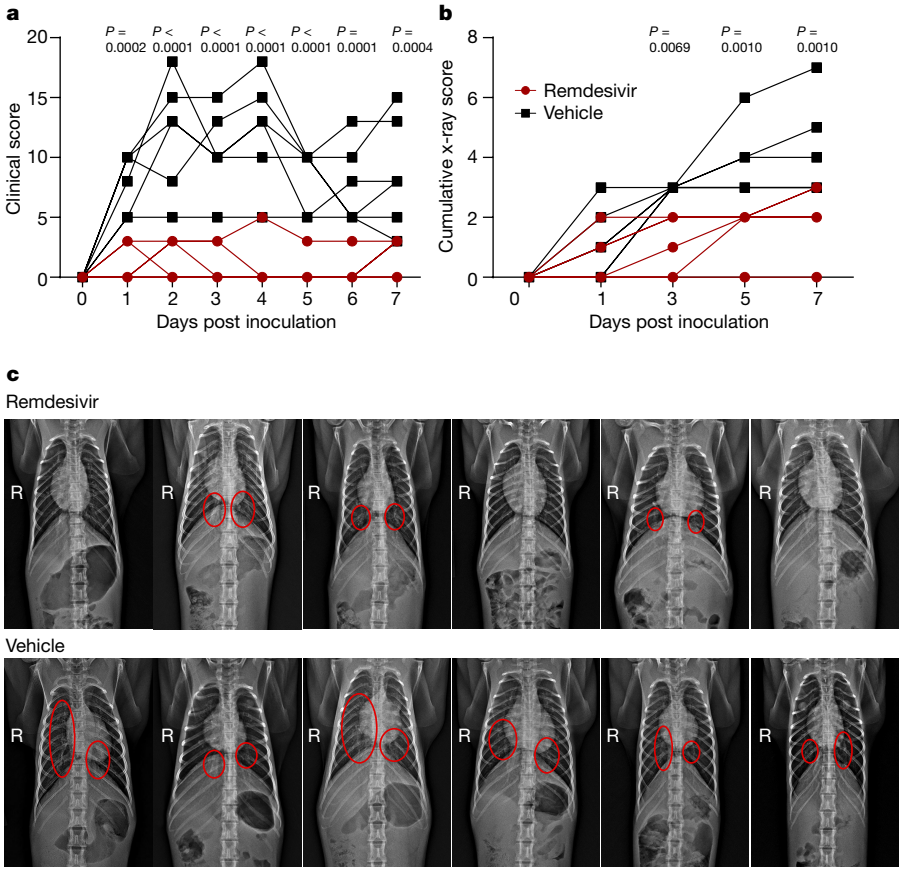


Fig. 1 | Reduced respiratory disease in rhesus macaques infected with SARS-CoV-2 and treated with remdesivir. a, Daily clinical scores for animals infected with SARS-CoV-2 and treated with remdesivir (red circles, $n = 6$) or vehicle solution (black squares, $n = 6$). **b**, Cumulative radiograph scores. Ventrodorsal and lateral radiographs were scored for the presence of pulmonary infiltrates by a clinical veterinarian according to a standard scoring system (0, normal; 1, mild interstitial pulmonary infiltrates; 2, moderate pulmonary infiltrates perhaps with partial cardiac border effacement and small areas of pulmonary consolidation; 3, severe interstitial infiltrates, large areas of pulmonary consolidation, alveolar patterns and air bronchograms). Individual lobes were scored and scores per animal per day were totalled and displayed. **c**, Ventrodorsal radiographs for each animal taken on 7 dpi. Areas of pulmonary infiltration are circled. Statistical analysis was performed using a two-way ANOVA with Sidak's multiple comparisons test.

the triphosphate of GS-441524, lung homogenate samples spiked with the triphosphate metabolite demonstrated rapid decay of the metabolite in this matrix (data not shown). GS-441524 levels were taken as a surrogate for tissue loading and suggest that the current dosing strategy delivered drug metabolites to the sites of SARS-CoV-2 replication in infected animals.

Lack of respiratory disease

After inoculation with SARS-CoV-2, the animals were assigned a daily clinical score based on a pre-established scoring sheet in a blinded fashion. Twelve hours after the first administration of remdesivir, clinical scores in remdesivir-treated animals were significantly lower

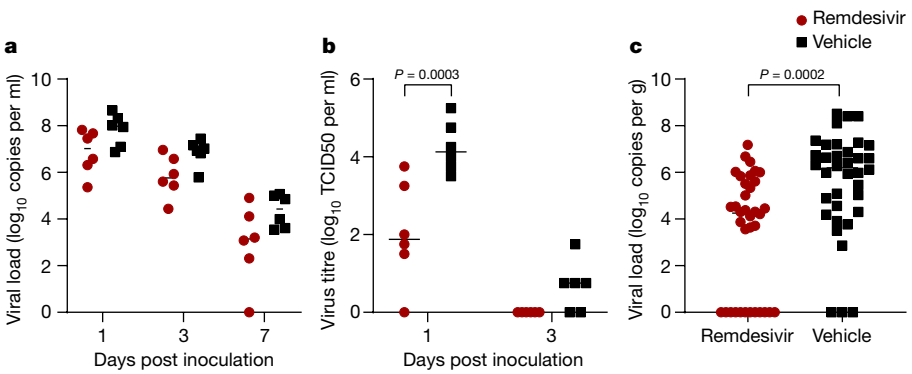


Fig. 2 | Viral loads and virus titres in BAL fluid and lung lobes. a, **b**, Viral loads (**a**) and infectious virus titres in BAL (**b**) collected from rhesus macaques infected with SARS-CoV-2 and treated with remdesivir ($n = 6$) or vehicle solution ($n = 6$). TCID50, 50% tissue culture infectious dose. Statistical analysis was performed using a two-way ANOVA with Sidak's multiple comparisons test. **c**, Viral loads in tissues collected from all six lung lobes at necropsy on 7 dpi from rhesus macaques infected with SARS-CoV-2 and treated with remdesivir ($n = 6$) or vehicle solution ($n = 6$). Statistical analysis was performed using an unpaired *t*-test. Centre bars, median.

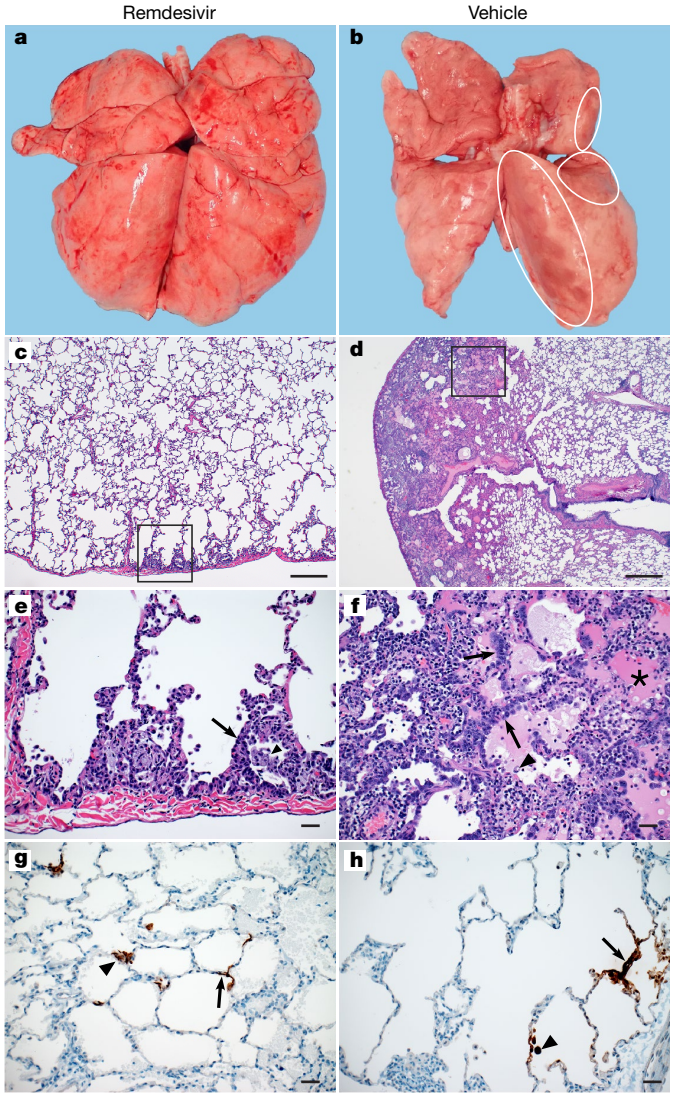


Fig. 3 | Changes to the lungs of rhesus macaques infected with SARS-CoV-2 and treated with remdesivir. Rhesus macaques infected with SARS-CoV-2 and treated with remdesivir (left, $n = 6$) or vehicle solution (right, $n = 6$) were euthanized on 7 dpi. **a**, **b**, Representative dorsal views of lungs from a remdesivir-treated animal (**a**) and a vehicle-treated animal with focally extensive areas of consolidation (**b**, circles). Histological analysis was performed on three sections from six lung lobes from each of the six animals per treatment group and representative images were chosen for **c**–**h**. **c**, Minimal subpleural interstitial pneumonia (box) observed in three out of six remdesivir-treated animals. **d**, Moderate subpleural interstitial pneumonia with oedema (box) observed in five of six vehicle-treated animals. **e**, Boxed area from **c** with alveoli lined by type II pneumocytes (arrow) and alveolar spaces containing foamy macrophages (arrowhead). **f**, Boxed area from **d** with pulmonary interstitium expanded by oedema and moderate numbers of macrophages and neutrophils. Alveoli are lined by type II pneumocytes (arrows). Alveolar spaces are filled with oedema (asterisk) and small numbers of pulmonary macrophages (arrowhead). **g**, Viral antigen in type I pneumocytes (arrow) and type II pneumocytes (arrowhead) of a remdesivir-treated animal. **h**, Viral antigen in type I pneumocytes (arrow) and macrophage (arrowhead) of a vehicle-treated animal. Scale bars: **c**, **d**, 200 µm; **e**–**h**, 20 µm.

than in control animals receiving vehicle solution. This difference in clinical score was maintained throughout the study (Fig. 1a). Only one of the six remdesivir-treated animals showed mild dyspnea, whereas tachypnea and dyspnea were observed in all vehicle-treated controls (Extended Data Table 1). Radiographic pulmonary infiltrates are one of the hallmarks of COVID-19 in humans. Radiographs taken on 0, 1, 3, 5, and 7 dpi showed significantly less lung lobe involvement and less severe pulmonary infiltration in animals treated with remdesivir than in those treated with vehicle (Fig. 1b, c).

Reduced virus replication in lungs

On 1, 3 and 7 dpi, bronchoalveolar lavage (BAL) was performed as an indicator of virus replication in the lower respiratory tract. Although viral loads in BAL were reduced in animals treated with remdesivir, this difference was not statistically significant (Fig. 2a). However, 12 h after the first remdesivir treatment was administered, the infectious virus titre in BAL was about 100-fold lower in remdesivir-treated animals than in controls. By 3 dpi, infectious virus could no longer be detected in BAL from remdesivir-treated animals, whereas virus was still detected in BAL from four out of six control animals (Fig. 2b). Despite this reduction in virus replication in the lower respiratory tract, there was no reduction in viral load or infectious virus titre in nose, throat or rectal swabs collected from remdesivir-treated animals, except for a significant difference in virus titre in throat swabs collected on 1 dpi and in viral loads in throat swabs collected on 4 dpi (Extended Data Fig. 2).

All animals were euthanized on 7 dpi. Tissue samples were collected from each lung lobe to compare virus replication between remdesivir-treated and vehicle-treated animals. In 10 out of 36 lung lobe samples collected from remdesivir-treated animals, viral RNA could not be detected, whereas this was the case in only 3 out of 36 lung lobes collected from control animals. In general, comparison across individual lung lobes in the two groups showed a lower geometric mean of viral RNA in the remdesivir-treated group (Extended Data Fig. 3a). Together, these data show that the viral load was significantly lower in lungs from remdesivir-treated animals than in those from vehicle-treated controls (Fig. 2c). Virus could be isolated from lung lobes of five out of six vehicle-treated control animals, but not from any of the lung tissue collected from remdesivir-treated animals. Although quantitative PCR with reverse transcription (qRT–PCR) showed that fewer tissues from other positions in the respiratory tract were positive for viral RNA in remdesivir-treated animals than in controls, these differences were not statistically significant (Extended Data Fig. 3b).

Reduced pneumonia

At necropsy on 7 dpi, lungs were assessed grossly for lesions. Gross lung lesions were observed in one out of six remdesivir-treated animals. By contrast, all six vehicle-treated control animals had visible lesions, resulting in a statistically significant difference in the area of the lungs affected by lesions (Fig. 3a, b, Extended Data Fig. 4a, b). This difference was also evident when calculating the lung weight-to-bodyweight ratio as an indicator of pneumonia; this ratio was significantly lower in remdesivir-treated than in vehicle-treated animals (Extended Data Fig. 4c). Histologically, remdesivir-treated animals had fewer and less severe lesions than did vehicle-treated controls. Histological lung lesions were absent in three out of six remdesivir-treated animals; the three remaining animals developed minimal pulmonary pathology. Lesions in these animals were characterized as widely separated, minimal, interstitial pneumonia frequently located in subpleural spaces (Fig. 3c, e). Five out six vehicle-treated animals developed multifocal, mild-to-moderate, interstitial pneumonia (Fig. 3d, f). We detected viral antigen in small numbers of type I and type II pneumocytes and alveolar macrophages in all animals, regardless of treatment (Fig. 3g, h).

Absence of resistance mutations

We successfully carried out deep sequencing on samples from all remdesivir-treated animals and vehicle-treated controls. Known mutations in the RNA-dependent RNA polymerase that confer resistance to

remdesivir in coronaviruses[11] were not detected in any of the samples tested (Supplementary Table 1).

Discussion

Remdesivir is, to our knowledge, the first antiviral treatment to show proven efficacy against SARS-CoV-2 in an animal model of COVID-19. Treatment of rhesus macaques infected with SARS-CoV-2 with remdesivir reduced clinical disease and damage to the lungs. The remdesivir dosing used in rhesus macaques is equivalent to that used in humans; however, owing to the acute nature of the disease in rhesus macaques, it is hard to directly translate the timing of treatment used to corresponding disease stages in humans. In our study, treatment was administered close to the peak of virus replication in the lungs as indicated by viral loads in bronchoalveolar lavages and the first effects of treatment on clinical signs and virus replication were observed within 12 h. The efficacy of direct-acting antivirals against acute viral respiratory tract infections typically decreases with delays in treatment initation[12]. Thus, remdesivir treatment should be initiated as early as possible in patients with COVID-19 to achieve the maximum treatment effect.

Despite the lack of obvious respiratory signs and reduced virus replication in the lungs of remdesivir-treated animals, there was no reduction in virus shedding. This finding is very important for patient management, where a clinical improvement should not be interpreted as a lack of infectiousness. Although we have shown that remdesivir metabolites are found in the lower respiratory tract, drug levels in the upper respiratory tract have not been characterized and novel formulations with alternative routes of drug delivery should be considered to improve distribution to the upper respiratory tract, thereby reducing shedding and the potential transmission risk. However, as severe COVID-19 disease results from virus infection of the lungs, this organ is the main target of remdesivir treatment. The bioavailability and protective effect of remdesivir in the lungs of infected rhesus macaques support treatment of COVID-19 patients with remdesivir. Remdesivir treatment did not result in a clinical improvement in one clinical trial with patients with severe COVID-19[13]; however, another clinical trial that involved more patients showed that remdesivir treatment resulted in a shorter time to improvement than in patients who received standard care only[14]. Our findings in rhesus macaques indicate that remdesivir treatment should be considered as early as clinically possible to prevent progression to pneumonia in patients with COVID-19.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2423-5.

1. Lo, M. K. et al. GS-5734 and its parent nucleoside analog inhibit Filo-, Pneumo-, and Paramyxoviruses. *Sci. Rep*. **7**, 43395 (2017).
2. Sheahan, T. P. et al. Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses. *Sci. Transl. Med*. **9**, eaal3653 (2017).
3. US National Library of Medicine *Clinical Trials.gov* (accessed 15 May 2020); https://clinicaltrials.gov/ct2/results?cond=&term=remdesivir&cntry=&state=&city=&dist=
4. US Food and Drug Administration. *Coronavirus (COVID-19) Update: FDA Issues Emergency Use Authorization for Potential COVID-19 Treatment* https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-issues-emergency-use-authorization-potential-covid-19-treatment (2020).
5. de Wit, E. et al. Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection. *Proc. Natl Acad. Sci. USA* **117**, 6771–6776 (2020).
6. Sheahan, T. P. et al. Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun*. **11**, 222 (2020).
7. Choy, K. T. et al. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro. *Antiviral Res*. **178**, 104786 (2020).
8. Wang, M. et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res*. **30**, 269–271 (2020).
9. Munster, V. J. et al. Respiratory disease in rhesus macaques inoculated with SARS-CoV-2. *Nature* https://doi.org/10.1038/s41586-020-2324-7 (2020).
10. Warren, T. K. et al. Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* **531**, 381–385 (2016).
11. Agostini, M. L. et al. Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. *MBio* **9**, e00221-18 (2018).
12. Sheahan, T. P. et al. An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Sci. Transl. Med*. **12**, eabb5883 (2020).
13. Wang, Y. et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* **395**, 1569–1578 (2020).
14. *Fact Sheet for Health Care Providers Emergency Use Authorization (EUA) of Veklury (remdesivir)* https://www.gilead.com/-/media/files/pdfs/remdesivir/eua-fact-sheet-for-hcps_01may2020.pdf (Gilead Sciences, 2020).

Methods

Ethics and biosafety statement

All animal experiments were approved by the Institutional Animal Care and Use Committee of Rocky Mountain Laboratories, NIH and carried out by certified staff in an Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International-accredited facility, according to the institution's guidelines for animal use, following the guidelines and basic principles in the NIH Guide for the Care and Use of Laboratory Animals, the Animal Welfare Act, United States Department of Agriculture and the United States Public Health Service Policy on Humane Care and Use of Laboratory Animals. Rhesus macaques were housed in adjacent individual primate cages, allowing social interactions, in a climate-controlled room with a fixed light–dark cycle (12 h light–12 h dark). Animals were monitored at least twice daily throughout the experiment. Commercial monkey chow, treats, and fruit were provided twice daily by trained personnel. Water was available ad libitum. Environmental enrichment consisted of a variety of human interactions, manipulanda, commercial toys, videos, and music. The Institutional Biosafety Committee (IBC) approved work with infectious SARS-CoV-2 strains under BSL3 conditions. Sample inactivation was performed according to IBC-approved standard operating procedures for removal of specimens from high containment.

Study design

To evaluate the effect of remdesivir treatment on SARS-CoV-2 disease outcome, we used the recently established rhesus macaque model of SARS-CoV-2 infection that results in transient lower respiratory tract disease[9]. Since this is a model with little prior data, it was not possible to perform a power analysis to determine group size. The sample size was therefore based on experience with other nonhuman primate models of respiratory disease, mainly a rhesus macaque model of MERS-CoV where $n = 6$ yielded statistical significance. Twelve animals were randomly assigned to two groups and inoculated as described previously with a total dose of 2.6×10^6 TCID50 (50% tissue culture infectious dose) of SARS-CoV-2 strain nCoV-WA1-2020 via intranasal, oral, ocular and intratracheal routes. The efficacy of therapeutic remdesivir treatment was tested in two groups of six adult rhesus macaques (three males and three females each; 3.6–5.7 kg). Owing to the acute nature of the SARS-CoV-2 model in rhesus macaques, therapeutic treatment was initiated 12 h after inoculation with SARS-CoV-2 and continued once daily for 6 dpi. One group of rhesus macaques was treated with a loading dose of 10 mg/kg remdesivir, followed by a daily maintenance dose of 5 mg/kg. The other group of six animals served as infected controls and were administered an equal dose volume (that is, 2 ml/kg loading dose and 1 ml/kg thereafter) of vehicle solution (12% sulfobutylether-β-cyclodextrin in water and hydrochloric acid, pH 3.5) according to the same treatment schedule. This dosing scheme in rhesus macaques mimics the daily dosing tested in clinical studies involving patients with COVID-19 and results in similar systemic drug exposure. Treatment was delivered as an intravenous bolus injection (total dose delivered over approximately 5 min) administered alternatingly in the left or right cephalic or saphenous veins. Although the remdesivir treatment course used here in rhesus macaques is shorter than the standard 10-day course in patients, this shorter treatment course was chosen to enable assessment of lung pathology at a time after inoculation when pulmonary infiltrates and interstitial pneumonia would still be present. Recent data from clinical trials have shown that a 5-day treatment course has a similar clinical benefit to a 10-day treatment course in patients with COVID-19[14].

The animals were observed twice daily for clinical signs of disease using a standardized scoring sheet as described previously[9]; the same person, who was blinded to the group assignment of the animals, assessed the animals throughout the study. The predetermined endpoint for this experiment was 7 dpi. Nose, throat and rectal swabs were collected daily during treatment administration. Clinical exams were performed on 0, 1, 3, 5, and 7 dpi on anaesthetized animals. On exam days, clinical parameters such as bodyweight, temperature, pulse oximetry, blood pressure and respiration rate were collected, as well as dorsoventral and lateral chest radiographs. Radiographs were analysed by a clinical veterinarian blinded to the group assignment of the animals. On 1, 3 and 7 dpi a BAL was performed using 10 ml sterile saline. After death on 7 dpi, necropsies were performed on the animals. The percentage of gross lung lesions were scored by a board-certified veterinary pathologist blinded to the group assignment of the animals and samples of the following tissues were collected: cervical lymph node, conjunctiva, nasal mucosa, oropharynx, tonsil, trachea, all lung lobes, mediastinal lymph node, right and left bronchus, heart, liver, spleen, kidney, stomach, duodenum, jejunum, ileum, caecum, colon, and urinary bladder. Histopathological analysis of tissue slides was performed by a board-certified veterinary pathologist blinded to the group assignment of the animals.

Virus and cells

SARS-CoV-2 isolate nCoV-WA1-2020 (MN985325.1)[15] (Vero passage 3) was kindly provided by the Centers for Disease Control (CDC) and propagated once in Vero E6 cells in Dulbecco's modified Eagle's medium (DMEM, Sigma) supplemented with 2% fetal bovine serum (Gibco), 1 mM L-glutamine (Gibco), 50 U/ml penicillin and 50 μg/ml streptomycin (Gibco) (virus isolation medium). The virus stock used was 100% identical to the initial deposited GenBank sequence (MN985325.1) and no contaminants were detected. VeroE6 cells were maintained in DMEM supplemented with 10% fetal calf serum, 1 mM L-glutamine, 50 U/ml penicillin and 50 μg/ml streptomycin.

Remdesivir (GS-5734)

Remdesivir (RDV; GS-5734) was manufactured at Gilead Sciences by the Department of Process Chemistry (Alberta, Canada) under Good Manufacturing Practice (GMP) conditions. Batch number 5734-BC-1P was solubilized in 12% sulfobutylether-β-cyclodextrin in water and matching vehicle solution was provided to NIH.

Liquid chromatography mass spectrometry (LC–MS)

Tributylamine was purchased from Millipore Sigma. LC–MS grade water, acetone, methanol, isopropanol and acetic acid were purchased through Fisher Scientific. All synthetic standards for molecular analysis were provided by Gilead Sciences Inc. Serum and cleared lung homogenates were gamma-irradiated (2 MRad) to inactivate infectious virus potentially present in these samples before analysis. Samples were prepared for small molecule analysis by diluting a 50-μl aliquot of either serum or clarified lung homogenate with 950 μl of 50% acetone, 35% methanol, 15% water (v/v) on ice. Samples were incubated at room temperature for 15 min and then centrifuged at 16,000*g* for 5 min. The clarified supernatants (850 μl) were recovered and taken to dryness in a Savant DNA120 SpeedVac concentrator (Thermo Fisher). Samples were resuspended in 100 μl of 50% methanol, 50% water (v/v) and centrifuged as before. The supernatant was taken to a sample vial for LC–MS analysis. Samples were separated using an ion-pairing liquid chromatography strategy on a Sciex ExionLC AC system. Samples were injected onto a Waters Atlantis T3 column (100 Å, 3 μm, 3 mm × 100 mm) and eluted using a binary gradient from 5 mM tributylamine, 5 mM acetic acid in 2% isopropanol, 5% methanol, 93% water (v/v) to 100% isopropanol over 5.5 min. Analytes were measured using a Sciex 5500 QTRAP mass spectrometer in negative mode. Multiple reaction monitoring was performed using two signal pairs for each analyte and signal fidelity was confirmed by collecting triggered product ion spectra and comparing back to spectra of synthetically pure standards.

All analytes were quantified against an eight-point calibration curve of the respective synthetic standard prepared in the target matrix (that is, serum or cleared lung homogenate) and processed in the same

manner as experimental samples. Limit of quantification (LOQ) was approximated at a signal to noise of 10. The LOQs for the measured molecules in each matrix were 5 nM for GS-441524 in both lung homogenate and serum, 1 nM for GS-704277 in both lung homogenate and serum and 0.08 nM for GS-5734 in serum. Instability of GS-5734 and the tri-phosphorylated nucleotide metabolite in the lung homogenate during tissue lysis prevented detection of these metabolites in the lung tissue.

Quantitative PCR

RNA was extracted from swabs and BAL using the QiaAmp Viral RNA kit (Qiagen) according to the manufacturer's instructions. Tissues (30 mg) were homogenized in RLT buffer and RNA was extracted using the RNeasy kit (Qiagen) according to the manufacturer's instructions. For detection of viral RNA, 5 µl RNA was used in a one-step real-time RT–PCR E assay[16] using the Rotor-Gene probe kit (Qiagen) according to the manufacturer's instructions. In each run, standard dilutions of RNA standards counted by droplet digital PCR were run in parallel, to calculate copy numbers in the samples.

Virus titration

Virus titrations were performed by end-point titration in Vero E6 cells. Tissue was homogenized in 1 ml DMEM using a TissueLyser (Qiagen). Cells were inoculated with tenfold serial dilutions of swab and BAL samples. Virus isolation was performed on lung tissues by homogenizing the tissue in 1 ml DMEM and inoculating Vero E6 cells in a 24-well plate with 250 µl cleared homogenate and a 1:10 dilution thereof. One hour after inoculation of cells, the inoculum was removed and replaced with 100 µl (virus titration) or 500 µl virus isolation medium. Six days after inoculation, CPE was scored and the TCID50 was calculated.

Histopathology and immunohistochemistry

Histopathology and immunohistochemistry were performed on rhesus macaque tissues. After fixation for a minimum of 7 days in 10% neutral-buffered formalin and embedding in paraffin, tissue sections were stained with haematoxylin and eosin (H&E). To detect SARS-CoV-2 antigen, immunohistochemistry was performed using a custom-made rabbit antiserum against SARS-CoV-2 N at a 1:1,000 dilution. Stained slides were analysed by a board-certified veterinary pathologist.

Next generation sequencing of viral RNA

Viral RNA was extracted as described above. cDNAs were prepared as described, with minor modifications[17]. In brief, 3–12 µl of extracted RNA was depleted of rRNA using Ribo-Zero Gold H/M/R (Illumina) and then reverse-transcribed using random hexamers and SuperScript IV (ThermoFisher Scientific). Following RNaseH treatment, second strand synthesis was performed using Klenow fragment (New England Biolabs) and resulting double-stranded cDNAs were treated with a combined mixture of RiboShredder RNase Blend (Lucigen) and RNase, DNase-free, high conc (Roche Diagnostics, Indianapolis, IN) and then purified using Ampure XP bead purification (Beckman Coulter). Kapa's HyperPlus library preparation kit (Roche Sequencing Solutions) was used to prepare sequencing libraries from the double-stranded cDNAs. To facilitate multiplexing, adaptor ligation was performed with KAPA Unique Dual-Indexed Adapters and samples were enriched for adaptor-ligated product using KAPA HiFi HotStart Ready mix and seven PCR amplification cycles, according to the manufacturer's manual. Pools consisting of eight sample libraries were used for hybrid-capture virus enrichment using myBaits Expert Virus SARS-CoV-2 panel and following the manufacturer's manual, version 4.01, with 14 cycles of post-capture PCR amplification (Arbor Biosciences). Purified, enriched libraries were quantified using Kapa Library Quantification kit (Roche Sequencing Solutions) and sequenced as 2×150-base pair reads on the Illumina NextSeq 550 instrument (Illumina).

Raw fastq reads were trimmed of Illumina adaptor sequences using cutadapt version 1.12[18] and then trimmed and filtered for quality using the FASTX-Toolkit (Hannon Lab). Remaining reads were mapped to the SARS-CoV-2 2019-nCoV/USA-WA1/2020 genome (MN985325.1) using Bowtie2 version 2.2.9[19] with parameters --local --no-mixed -X1500. PCR duplicates were removed using picard MarkDuplicates (Broad Institute) and variants were called using GATK HaplotypeCaller version 4.1.2.0[20] with parameter -ploidy 2. Variants were filtered for QUAL >1000 and DP >20 using bcftools.

Statistical analysis

Statistical analyses were performed using GraphPad Prism software version 8.2.1.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data included in this manuscript have been deposited in Figshare (https://doi.org/10.35092/yhjc.12111570). Sequences have been deposited in NCBI under BioProject accession number PRJNA632475.

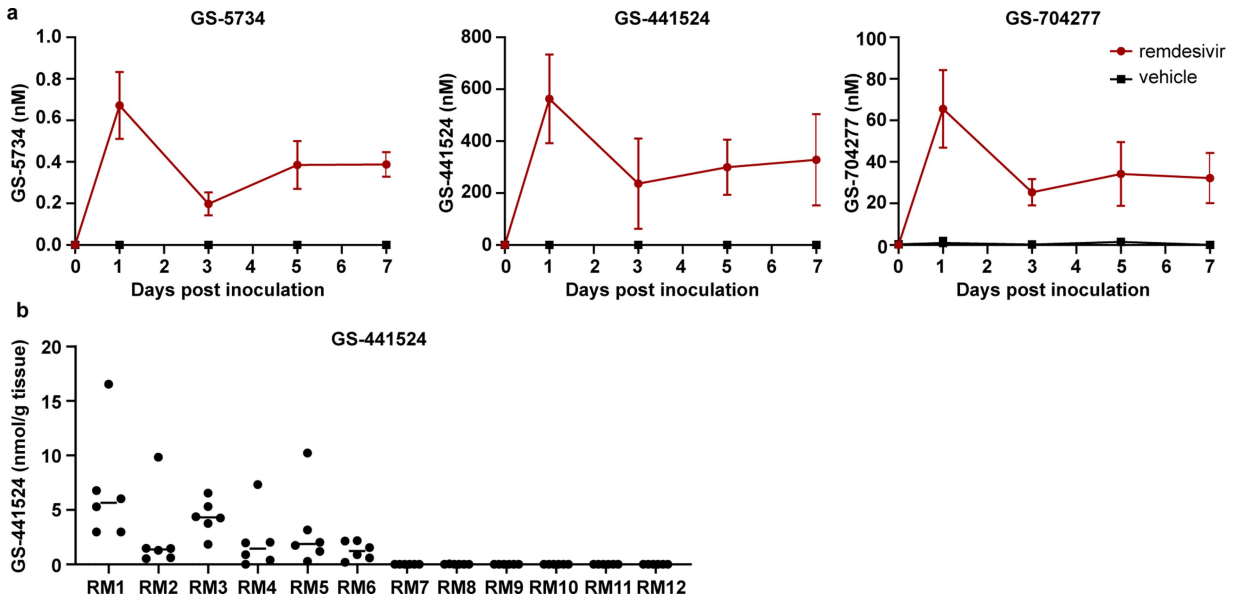
15. Harcourt, J. et al. Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. *Emerg. Infect. Dis*. **26**, 1266–1273 (2020).
16. Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill*. **25**, 2000045 (2020).
17. Briese, T. et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio* **6**, e01491-15 (2015).
18. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
19. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
20. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).

Acknowledgements We thank E. Bunyan for preparing remdesivir; D. Babusis for providing synthetic standards for molecular analysis; A. Mora for preparing figures; T. Thomas, R. Rosenke and D. Long for assistance with histology; M. Holbrook and T. Bushmaker for technical assistance; and RMVB staff for animal care. This study was supported by the Intramural Research Program of NIAID, NIH.

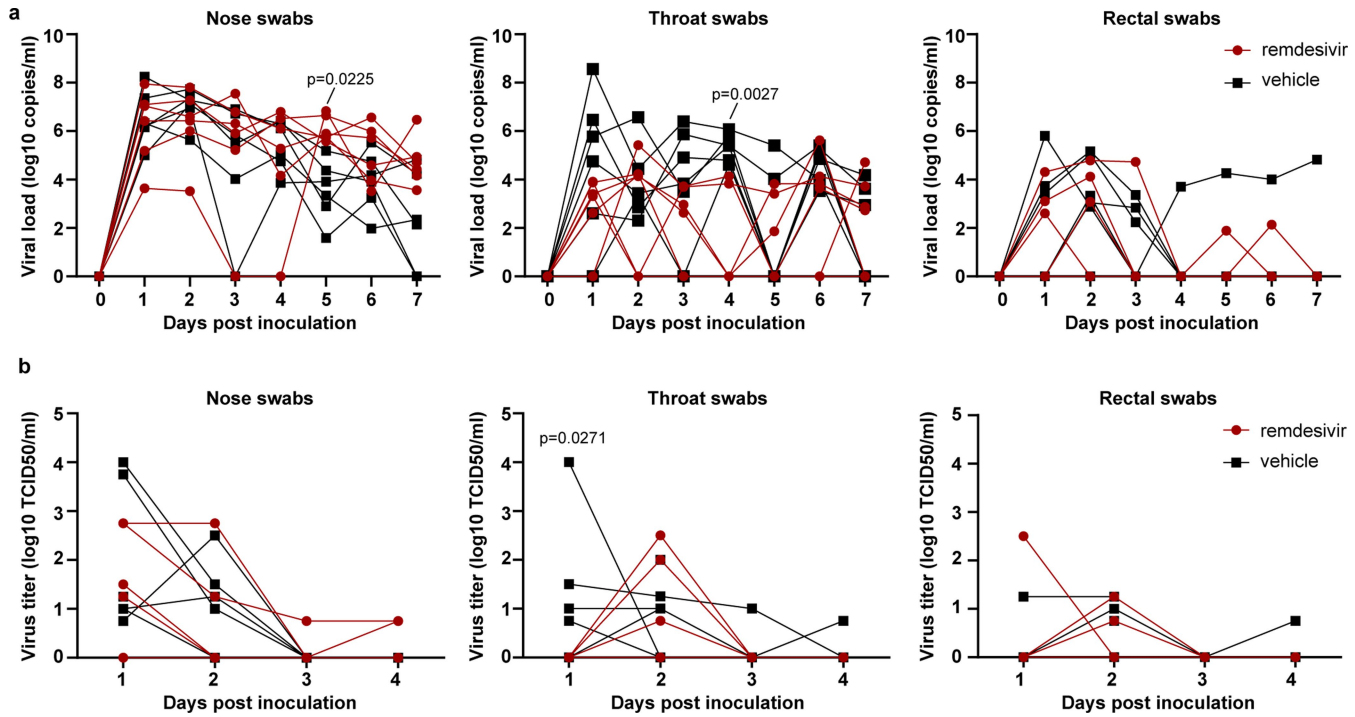
Author contributions D.P.P., T.C., V.J.M. and E.d.W. designed the study; B.N.W., F.F., B.S., K.M.-W., J.S., N.v.D., I.L., C.K.Y., L.P.-P., A.O., J.L., P.W.H., G.S., S.A., K.B., C.M., D.P.S., V.L.M. and E.d.W. acquired and analysed the data; B.N.W., F.F., B.S., D.P.P., N.v.D., C.K.Y., A.O., J.L., P.W.H., G.S., C.M.B., S.A., K.B., T.C., C.M., D.P.S., V.J.M. and E.d.W. interpreted the data; and E.d.W. wrote the manuscript. All authors approved the submitted version of the manuscript.

Competing interests D.P.P. and T.C. are employees of Gilead Sciences and own company stock. The remaining authors report no competing interests.

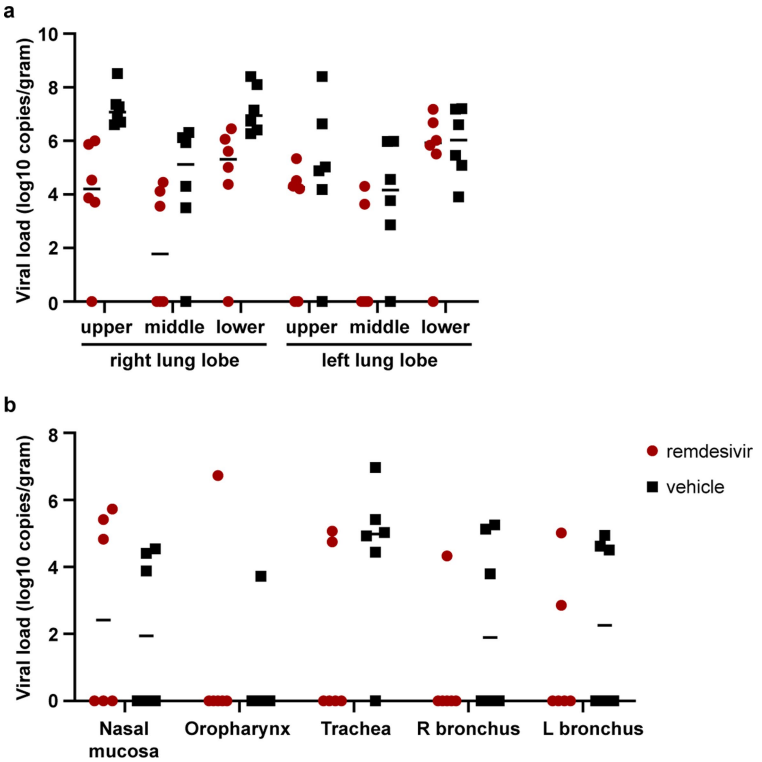
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2423-5.
Correspondence and requests for materials should be addressed to E.d.W.
Peer review information *Nature* thanks Wolfgang Baumgärtner, Stanley Pearlman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



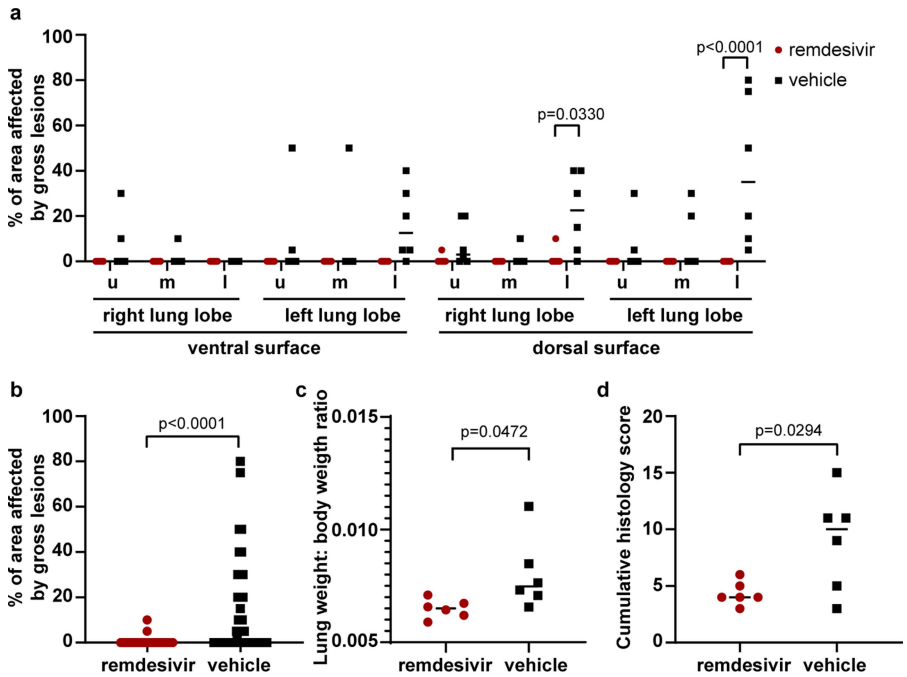
Extended Data Fig. 1 | Concentration of remdesivir prodrug and metabolites measured in serum and lung homogenates of rhesus macaques infected with SARS-CoV-2. Two groups of six rhesus macaques were inoculated with SARS-CoV-2 strain nCoV-WA1-2020. Twelve hours post inoculation, one group was administered 10mg/kg intravenous remdesivir and the other group was treated with an equal volume of vehicle solution (2 ml/kg). Treatment was continued 12hrs after the first treatment, and every 24 h thereafter with a dose of 5 mg/kg remdesivir or equal volume of vehicle solution (1 ml/kg). **a**, Serum concentration of remdesivir prodrug GS-5734, the dephosphorylated nucleoside product GS-441524 and the intermediate alanine metabolite GS-704277 over time as measured by LCMS for all animals ($n = 12$) in the study. Mean and standard deviation are shown. **b**, Concentration of GS-441524 homogenized lung tissue collected from all six lung lobes from each animal ($n = 12$) on 7 dpi, 24 h after the last remdesivir treatment was administered. Each dot represents the concentration of GS-441524 in one lung lobe. The centre bar represents the median.



Extended Data Fig. 2 | Viral loads and virus titres in swabs collected from rhesus macaques infected with SARS-CoV-2 and treated with remdesivir. **a**, Viral loads; **b**, infectious virus titers in nose, throat and rectal swabs collected daily from animals treated with remdesivir ($n = 6$) or vehicle solution ($n = 6$). Statistical analysis was performed using a 2-way ANOVA with Sidak's multiple comparisons test.



Extended Data Fig. 3 | Viral loads in tissues collected from the respiratory tract on 7 dpi. a, Viral loads in all six lung lobes collected from rhesus macaques infected with SARS-CoV-2 and treated with remdesivir ($n = 6$) or vehicle solution ($n = 6$), stratified per lung lobe. **b**, Viral loads in other tissues collected throughout the respiratory tract on 7 dpi. The centre bar represents the median.



Extended Data Fig. 4 | Pathological changes in lungs of rhesus macaques infected with SARS-CoV-2 and treated with remdesivir. Rhesus macaques infected with SARS-CoV-2 and treated with remdesivir ($n = 6$) or vehicle solution ($n = 6$) were euthanized on 7 dpi. **a**, The area of each individual lung lobe affected by gross lesions as scored by a veterinary pathologist blinded to group assignment of the animals. **b**, All data from panel a combined. **c**, Lung weight: bodyweight ratio as an indicator of pulmonary oedema. **d**, Cumulative histology score. Each lung lobe was scored for the presence of histological lung lesions on a predetermined scale (0–4); these values were combined per animal and graphed. Data in panel a were analysed using a 2-way ANOVA with Sidak's multiple comparisons test; data in **b**–**d** were analysed using a two-tailed, unpaired *t*-test. The centre bar represents the median.

Extended Data Table 1 | Clinical and pathological observations in rhesus macaques inoculated with SARS-CoV-2 and treated with remdesivir

Treatment	Animal	Clinical observations	Observations at necropsy
Remdesivir	RM1	Slightly decreased appetite	Mediastinal lymph nodes enlarged
	RM2	Slightly decreased appetite	None
	RM3	Slightly decreased appetite, pale appearance	Mediastinal lymph nodes enlarged
	RM4	Slightly decreased appetite, slightly dehydrated	Mediastinal lymph nodes enlarged
	RM5	Slightly decreased appetite	Mediastinal lymph nodes enlarged
	RM6	Mild dyspnea, pale appearance	Gross lung lesions; mediastinal lymph nodes enlarged
Vehicle solution	RM7	Piloerection, hunched posture, tachypnea, dyspnea, decreased appetite	Gross lung lesions; mediastinal lymph nodes enlarged; focal hemorrhage in colon
	RM8	Piloerection, hunched posture, tachypnea, dyspnea, decreased appetite	Gross lung lesions; mediastinal lymph nodes enlarged
	RM9	Piloerection, hunched posture, tachypnea, dyspnea, decreased appetite	Gross lung lesions; mediastinal lymph nodes enlarged
	RM10	Tachypnea, dyspnea, pale appearance, slightly dehydrated	Gross lung lesions; mediastinal lymph nodes enlarged
	RM11	Piloerection, tachypnea, dyspnea, decreased appetite, pale appearance	Gross lung lesions; mediastinal lymph nodes enlarged
	RM12	Piloerection, tachypnea, dyspnea, decreased appetite	Gross lung lesions; mediastinal lymph nodes enlarged; ~5ml fluid in peritoneum

nature research

Corresponding author(s): Emmie de Wit

Last updated by author(s): May 15, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequence analysis software used: Cutadapt version 1.12; FASTX-Toolkit; Bowtie2 version 2.2.9; picard MarkDuplicates; GATK HaplotypeCaller version 4.1.2.0; bcftools; GATK version 3 DepthOfCoverage tool
Data analysis	Data were analyzed using Graphpad Prism 8.2.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data have been deposited in Figshare: https://doi.org/10.1101/2020.04.15.043166
Sequences have been deposited in NCBI, BioProject accession number PRJNA632475

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Since this is a model with little prior data, it was not possible to perform a power analysis. The sample size was based on experience with other nonhuman primate models of respiratory disease, mainly a rhesus macaque model of MERS-CoV where n=6 yielded statistical significance.
Data exclusions	No data were excluded.
Replication	Lung histology: for each animal (n=6 per group), 3 sections were evaluated from all 6 lung lobes. Radiographs: Three chest radiographs were taken from each animal at each clinical exam: right-lateral, left-lateral and ventro-dorsal; only the ventro-dorsal radiograph is shown.
Randomization	Animals were randomly assigned to the group administered remdesivir or vehicle solution.
Blinding	The following tasks were performed by researchers blinded to group assignment: daily clinical scoring; analysis of radiographs; histopathology

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
[ ]	[x] Antibodies
[ ]	[x] Eukaryotic cell lines
[x]	[ ] Palaeontology
[ ]	[x] Animals and other organisms
[x]	[ ] Human research participants
[x]	[ ] Clinical data

Methods

n/a	Involved in the study
[x]	[ ] ChIP-seq
[x]	[ ] Flow cytometry
[x]	[ ] MRI-based neuroimaging

Antibodies

Antibodies used	Custom-ordered anti-SARS-CoV-2 nucleocapsid antibody; generated in rabbits by GenScript. Since this is a custom order there is no catalog number for this antibody.
Validation	Validation of cross-reactivity of the SARS-CoV-2 custom antibody in IHC was done in-house by embedding SARS-CoV-2 infected Vero cells in histogel and producing and staining histology slides; this was then confirmed by staining known SARS-CoV-2-postive lung tissue as well as negative control tissue.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	VeroE6: Ralph Baric, University of North Carolina, Chapel Hill, USA
Authentication	Not authenticated in-house.
Mycoplasma contamination	Mycoplasma testing confirmed negative at regular intervals.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Rhesus macaques, Chinese origin, adult (4-6 years), 6 males, 6 females
Wild animals	No wild animals were used.

Field-collected samples	No samples were collected in the field.
Ethics oversight	All animal experiments were approved by the Institutional Animal Care and Use Committee of Rocky Mountain Laboratories, NIH and carried out by certified staff in an Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International accredited facility, according to the institution's guidelines for animal use, following the guidelines and basic principles in the NIH Guide for the Care and Use of Laboratory Animals, the Animal Welfare Act, United States Department of Agriculture and the United States Public Health Service Policy on Humane Care and Use of Laboratory Animals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Cancer SLC43A2 alters T cell methionine metabolism and histone methylation

https://doi.org/10.1038/s41586-020-2682-1

Received: 28 July 2019

Accepted: 29 May 2020

Published online: 2 September 2020

Check for updates

Yingjie Bian[1,2,16], Wei Li[1,2,16], Daniel M. Kremer[3], Peter Sajjakulnukit[3], Shasha Li[1,2,4], Joel Crespo[1,2], Zeribe C. Nwosu[3], Li Zhang[3], Arkadiusz Czerwonka[5], Anna Pawłowska[6], Houjun Xia[1,2], Jing Li[1,2], Peng Liao[1,2], Jiali Yu[1,2], Linda Vatan[1,2], Wojciech Szeliga[1,2], Shuang Wei[1,2], Sara Grove[1,2], J. Rebecca Liu[7], Karen McLean[7], Marcin Cieslik[4,8], Arul M. Chinnaiyan[8,9,10,11], Witold Zgodziński[12], Grzegorz Wallner[12], Iwona Wertel[6], Karolina Okła[6], Ilona Kryczek[1,2], Costas A. Lyssiotis[3,13,14,15] & Weiping Zou[1,2,8,14,15] ✉

Abnormal epigenetic patterns correlate with effector T cell malfunction in tumours[1–4], but the cause of this link is unknown. Here we show that tumour cells disrupt methionine metabolism in CD8$^+$ T cells, thereby lowering intracellular levels of methionine and the methyl donor *S*-adenosylmethionine (SAM) and resulting in loss of dimethylation at lysine 79 of histone H3 (H3K79me2). Loss of H3K79me2 led to low expression of STAT5 and impaired T cell immunity. Mechanistically, tumour cells avidly consumed methionine and outcompeted T cells for methionine by expressing high levels of the methionine transporter SLC43A2. Genetic and biochemical inhibition of tumour SLC43A2 restored H3K79me2 in T cells, thereby boosting spontaneous and checkpoint-induced tumour immunity. Moreover, methionine supplementation improved the expression of H3K79me2 and STAT5 in T cells, and this was accompanied by increased T cell immunity in tumour-bearing mice and patients with colon cancer. Clinically, tumour SLC43A2 correlated negatively with T cell histone methylation and functional gene signatures. Our results identify a mechanistic connection between methionine metabolism, histone patterns, and T cell immunity in the tumour microenvironment. Thus, cancer methionine consumption is an immune evasion mechanism, and targeting cancer methionine signalling may provide an immunotherapeutic approach.

Immune checkpoint blockade therapies have demonstrated unprecedented clinical efficacy in cancer treatment, but their application has been hindered by therapeutic resistance[5]. CD8$^+$ T cells mediate anti-tumour immunity. Unfortunately, tumour-infiltrating CD8$^+$ T cells are often dysfunctional (this is known as T cell exhaustion)[1,6]. Differentiation and activation of T cells are associated with changes to the epigenetic landscape at gene loci that encode effector molecules, including interferon (IFN) and granzyme B[4,7]. However, these dynamic epigenetic changes in T cells may be disrupted by tumour cells[2,4,8] via metabolic regulation in the tumour microenvironment[9–12]. Tumour-intrinsic mechanisms, and oncogenic signalling in particular, may contribute to abnormal tumour metabolism[13,14]. However, it is unclear whether amino acid metabolism can affect the T cell epigenetic landscape and in turn alter T cell function in tumours.

Tumour cells outcompete T cells for methionine

Exhausted T cells exhibit distinct histone profiles and limit tumour immunotherapy[2,3]. To test whether abnormal amino acid metabolism is related to alterations in histones and dysfunction in T cells, we cultured mouse CD8$^+$ T cells without individual amino acids. Omission of methionine resulted in the most marked T cell death and dysfunction, as shown by staining of cultured cells for the apoptosis marker annexin V and the effectors IFNγ and TNFα (Fig. 1a–c). Thus, access to methionine is critical for the survival and function of T cells.

Next, we tested whether tumour cells impaired CD8$^+$ T cell function by altering methionine levels. We cultured ID8 (Fig. 1d) and B16F10 (Fig. 1e) tumour cells with medium containing 20–100 μM methionine. Regardless of methionine concentrations, fresh medium had a minimal effect

[1]Department of Surgery, University of Michigan School of Medicine, Ann Arbor, MI, USA. [2]Center of Excellence for Cancer Immunology and Immunotherapy, University of Michigan Rogel Cancer Center, University of Michigan School of Medicine, Ann Arbor, MI, USA. [3]Department of Molecular and Integrative Physiology, University of Michigan Medical School, Ann Arbor, MI, USA. [4]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [5]Department of Virology and Immunology, Maria Curie-Skłodowska University, Lublin, Poland. [6]First Chair and Department of Oncological Gynecology and Gynecology, Medical University of Lublin, Lublin, Poland. [7]Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, MI, USA. [8]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [9]Department of Urology, University of Michigan, Ann Arbor, MI, USA. [10]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA. [11]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA. [12]Second Department of General, Gastrointestinal Surgery and Surgical Oncology of the Alimentary Tract, Medical University of Lublin, Lublin, Poland. [13]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA. [14]Graduate Program in Immunology, University of Michigan School of Medicine, Ann Arbor, MI, USA. [15]Graduate Program in Cancer Biology, University of Michigan School of Medicine, Ann Arbor, MI, USA. [16]These authors contributed equally: Yingjie Bian, Wei Li. ✉e-mail: wzou@med.umich.edu

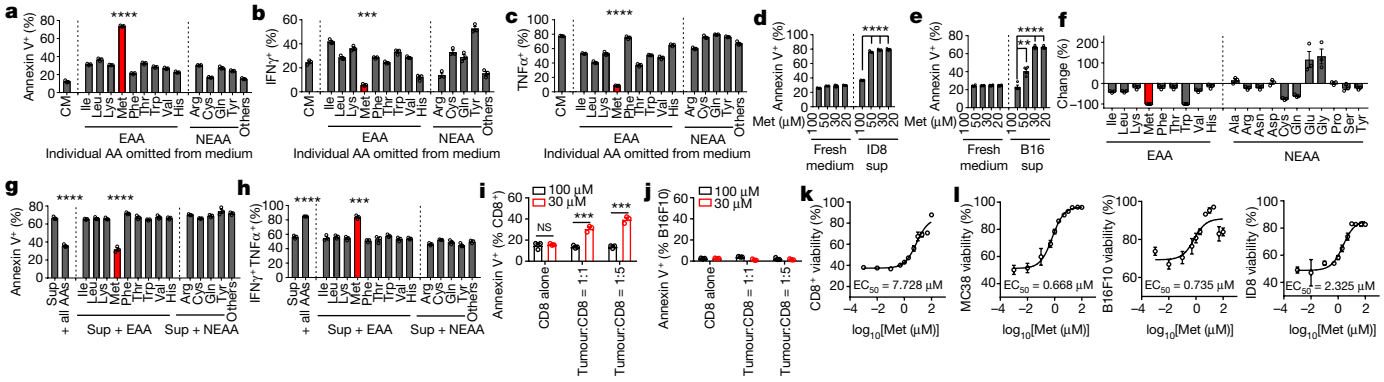


Fig. 1 | Tumour cells outcompete T cells for methionine to impair T cell function. a–**c**, Effects of amino acids on T cell apoptosis (**a**) and effector cytokines (**b**, **c**). Activated mouse CD8$^+$ T cells were cultured with complete medium (CM) or media from which individual amino acids (AAs) had been omitted for 36 h. EAA, essential amino acid; NEAA, non-essential amino acid. **d**, **e**, Effect of tumour cell culture supernatants on T cell apoptosis. CD8$^+$ T cells were cultured for 36 h with supernatants (sup) from cultured ID8 (**d**) or B16F10 (**e**) cells with varying concentrations of methionine (Met). **f**, Mass spectrometry (MS) detection of amino acid consumption by cultured tumour cells (bars show change from fresh medium). **g**, **h**, Effect of amino acid supplementation in tumour cell supernatant on T cell apoptosis (**g**) and cytokine production (**h**). CD8$^+$ T cells were cultured with B16F10 cell culture supernatant supplemented with amino acids for 36 h. **i**, **j**, Apoptosis of T cells (**i**) and tumour cells (**j**) induced by methionine competition. B16F10 cells and CD8$^+$ T cells were cultured at different ratios for 72 h in a Transwell system with 30 or 100 μM methionine. NS, not significant. **k**, **l**, Effect of methionine on viability of CD8$^+$ T cells (**k**) and tumour cells (**l**). Data are mean ± s.e.m. Sample sizes (*n*), *P* values, statistical tests and numbers of replications are listed in 'Statistics and reproducibility' (Methods).

on apoptosis in T cells (Fig. 1d, e). However, supernatant from cultured ID8 (Fig. 1d) or B16F10 cells (Fig. 1e) induced apoptosis in CD8$^+$ T cells when the original culture medium had contained less than 100 μM methionine. We obtained similar results when we cultured mouse CD8$^+$ T cells with the supernatant from MC38 or CT26 colon cancer cells (Extended Data Fig. 1a, b) or human CD8$^+$ T cells with the supernatant from A375 melanoma cells (Extended Data Fig. 1c). Moreover, supernatants from tumour cell cultures increased death and dysfunction in cultured ID8 tumour-infiltrating T cells (Extended Data Fig. 1d, e). Thus, tumour cells limit T cell access to methionine and impair T cell survival and function.

The physiological concentration of methionine in human serum is about 30 μM[15,16], but serum methionine was lower in patients with cancer than in healthy donors (Extended Data Fig. 1f, g). To evaluate methionine consumption within the physiological range, we cultured B16F10 cells with 30 μM methionine and analysed the abundance of amino acids in the supernatant. Tumour cells consumed the majority of amino acids, including methionine and tryptophan (Fig. 1f, Extended Data Fig. 1h). We subsequently cultured mouse CD8$^+$ T cells in B16F10 cell supernatants supplemented with individual amino acids. Among all amino acids, only methionine supplementation prevented T cell apoptosis and rescued the production of IFNγ and TNFα (Fig. 1g, h). We obtained similar results with human CD8$^+$ T cells (Extended Data Fig. 1i, j). Tumour glycolysis regulates T cell function[17], so we cultured T cells in tumour cell supernatant supplemented with glucose or methionine. Methionine, but not glucose, restored T cell survival and production of cytokines (Extended Data Fig. 1k–m). Moreover, simultaneous supplementation with glucose and methionine did not enhance this effect. Thus, consumption of methionine by tumour cells impairs T cell survival and function.

We cultured B16F10 and CD8$^+$ T cells in a Transwell system (Extended Data Fig. 1n). A high concentration of methionine (100 μM) had a minimal effect on apoptosis in tumour or CD8$^+$ T cells. However, a low concentration of methionine (30 μM) caused apoptosis in CD8$^+$ T cells (Fig. 1i), but not tumour cells (Fig. 1j). Then, we evaluated the half-maximal effective concentration (EC_{50}) of methionine to maintain CD8$^+$ T and tumour cell viabilities. Both mouse and human CD8$^+$ T cells were more sensitive than tumour cells to methionine deprivation, as shown by their different EC_{50} values (Fig. 1k, l, Extended Data Fig. 1o, p). Thus, tumour cells outcompete T cells for methionine, thereby impairing T cell survival and function.

Low methionine decreases H3K79me2 in T cells

To investigate the mechanism by which tumour cells affected CD8$^+$ T cells through methionine deprivation, we performed RNA sequencing (RNA-seq) on CD8$^+$ T cells cultured with fresh medium, B16F10 supernatant, and supernatant plus methionine (Extended Data Fig. 2a). Network grouping analysis revealed that pathways related to metabolism, function, and survival were affected by tumour supernatants (Extended Data Fig. 2b). Correspondingly, gene set enrichment analysis (GSEA) showed an enrichment in the T cell apoptosis signature and poor T cell receptor signalling in the presence of tumour supernatant (Fig. 2a), whereas methionine supplementation largely rescued this phenotype (Extended Data Fig. 2c). Moreover, one-carbon metabolic process and the methionine cycle were defective in CD8$^+$ T cells cultured with supernatant (Fig. 2b), and were restored by methionine addition (Extended Data Fig. 2d, e).

Next, we performed a metabolomics analysis of parallel CD8$^+$ T cells. We observed obvious metabolic changes in T cells cultured with tumour supernatants, and these too were rescued by methionine addition (Extended Data Fig. 2f). We specifically examined metabolites related to the one-carbon process and the methionine cycle (Fig. 2c, Extended Data Fig. 2g), and found that CD8$^+$ T cells cultured in tumour supernatant showed a marked decrease in intracellular methionine, SAM and *S*-adenosyl-homocysteine (SAH) (Fig. 2d–f). Supplementation with methionine restored intracellular methionine, SAM, and SAH (Fig. 2d–f), and induced a decrease in serine and L-cystathionine (Extended Data Fig. 2h, i). To test which metabolite was the key factor, we cultured CD8$^+$ T cells with tumour supernatant supplemented with methionine, SAM, SAH, or L-cystathionine. Supplementation with methionine or SAM prevented CD8$^+$ T cell apoptosis, and rescued the T cell cytokine profile (Fig. 2g, h).

Intracellular methionine is converted into SAM, the donor for epigenetic methylation[18,19]. Thus, we tested T cell histone marks and found that supernatants induced a marked decrease in H3K79me2, but not in other marks (Fig. 2i). Similar results were obtained in mouse CD8$^+$ T cells cultured with CT26 or MC38 supernatant and in human CD8$^+$ T cells

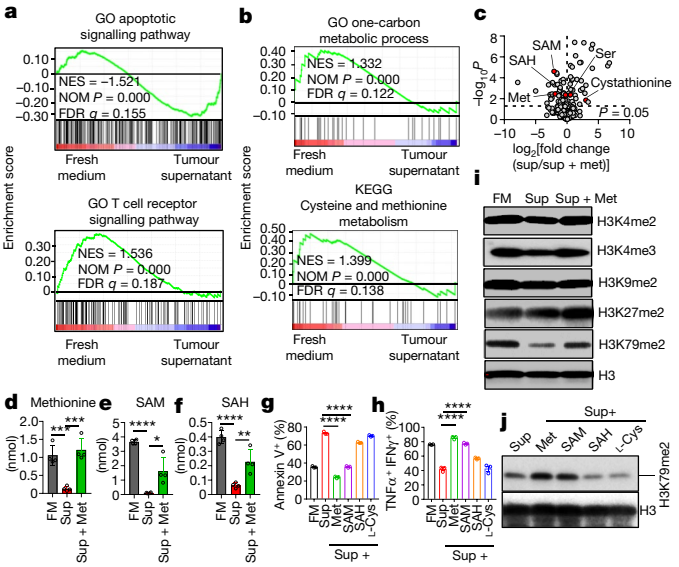


Fig. 2 | Tumour cells alter CD8⁺ T cell methionine metabolism to diminish H3K79me2. a, **b**, GSEA plot showing enriched apoptotic and TCR signalling pathways (**a**) and defective methionine metabolism signalling (**b**) in CD8⁺ T cells cultured in tumour cell culture supernatant. GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; NES, normalized enrichment score. **c**–**f**, Changes to the methionine metabolism pathway in CD8⁺ T cells cultured with fresh medium (FM), tumour cell supernatant, or supernatant supplemented with methionine. **c**, Volcano plot shows changes in metabolite levels between T cells cultured with supernatant and those cultured with supernatant and methionine. Intracellular methionine (**d**), SAM (**e**), and SAH concentrations (**f**) were detected by MS. **g**, **h**, Effect of metabolite supplementation on apoptosis (**g**) and cytokine production (**h**) in CD8⁺ T cells cultured in tumour cell supernatant. **i**, Effect of tumour supernatants on histone methylation in CD8⁺ T cells. **j**, Effect of methionine metabolite supplementation on H3K79 methylation in CD8⁺ T cells. Data are mean ± s.e.m. Sample sizes (*n*), *P* values, statistical tests and numbers of replications are listed in 'Statistics and reproducibility' (Methods).

cultured with A375 supernatants (Extended Data Fig. 2j, k). Moreover, the reduced H3K79me2 could be recovered by supplementation with methionine or SAM, but not SAH or L-cystathionine (Fig. 2j). Thus, methionine restriction by tumour cells reduces the methyl donor SAM and, in turn, impairs H3K79me2 in CD8⁺ T cells.

Loss of H3K79me2 impairs STAT5 expression

Disruptor of telomeric silencing 1-like (DOT1L) is the specific and sole methyltransferase for H3K79[20,21]. We cultured CD8⁺ T cells with EPZ004777, an inhibitor of DOT1L. EPZ004777 inhibited H3K79me2, induced CD8⁺ T cell apoptosis, and suppressed CD8⁺ T cell cytokine expression in a dose-dependent manner (Fig. 3a–c). To genetically explore the role of H3K79me2 in T cell function, we crossed conditional *Dot1l* allele (*Dot1l*$^{flox/flox}$, referred to here as *Dot1l*$^{+/+}$) mice[22] with CD4-Cre transgenic mice to delete *Dot1l* specifically in T cells (referred to here as *Dot1l*$^{-/-}$ mice) (Extended Data Fig. 3a). Deletion of *Dot1l* led to a loss of H3K79me2 in CD8⁺ T cells (Fig. 3d, Extended Data Fig. 3b) and resulted in increased apoptosis, especially upon activation (Fig. 3e). Moreover, intracellular cytokine staining showed impaired function in *Dot1l*$^{-/-}$ CD8⁺ T cells (Fig. 3f). RNA array analysis showed that *Dot1l*$^{-/-}$ CD8⁺ T cells (Extended Data Fig. 3c) were similar to T cells exposed to methionine deficiency (Extended Data Fig. 2b). For instance, like methionine-deficient T cells (Fig. 2a), *Dot1l*$^{-/-}$ T cells showed an enriched apoptotic gene signature and impaired T cell receptor functional gene signature (Extended Data Fig. 3d, e). The data suggest that T cell malfunctions caused by methionine deficiency or by impaired DOT1L-dependent histone methylation share a mechanism.

We assessed the role of T cell DOT1L in tumour immunity and found that MC38 tumours grew faster in *Dot1l*$^{-/-}$ mice than in *Dot1l*$^{+/+}$ mice (Fig. 3g, h). Correspondingly, *Dot1l*$^{-/-}$ mice showed an increase in CD8⁺ T cell apoptosis in tumour draining lymph nodes and tumour tissues (Fig. 3i), as well as a decrease in secretion of TNFα, IFNγ, and granzyme B from tumour-infiltrating CD8⁺ T cells (Extended Data Fig. 3f). In addition, PD-L1 blockade inhibited tumour growth in *Dot1l*$^{+/+}$ mice but not in *Dot1l*$^{-/-}$ mice (Extended Data Fig. 3g). We obtained similar results with B16F10 tumours (Extended Data Fig. 3h, i). To confirm that the positive role of methionine in T cells depends on DOT1L, we cultured *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ T cells with methionine supplementation in the presence of tumour supernatant. Methionine supplementation failed to protect *Dot1l*$^{-/-}$ T cells from apoptosis (Fig. 3j) or to rescue their impaired cytokine production (Fig. 3k). Thus, loss of DOT1L, which mediates H3K79me2, weakens anti-tumour T cell immunity.

We next investigated how loss of H3K79me2 results in T cell dysfunction. *Dot1l*$^{-/-}$ CD8⁺ T cells showed enrichment of an apoptosis gene signature (Fig. 3l, Extended Data Fig. 3c, d). The JAK–STAT pathway regulates T cell survival and effector function[23]. Among components of the JAK–STAT pathway, expression of *Stat5* was most strongly affected by H3K79me2 deficiency (Fig. 3m, Extended Data Fig. 3j); both total STAT5 and phosphorylated STAT5 (p-STAT5), but not other STATs, were reduced in *Dot1l*$^{-/-}$ T cells (Fig. 3n). In mouse CD8⁺ T cells cultured with B16F10 supernatant, *Stat5* transcripts (Extended Data Fig. 3k), total STAT5 and p-STAT5 (Fig. 3o) were all reduced. These effects were reversed by supplementation with methionine or SAM, but not SAH or L-cystathionine (Fig. 3o). Moreover, RNA-seq data from human CD8⁺ T cells treated with the DOT1L inhibitor SGC0946[24] showed reduced *STAT5* expression, enriched apoptotic gene signatures, and impaired T cell signalling (Extended Data Fig. 3l–n). Thus, tumour cells outcompete T cells for methionine, resulting in a reduction in H3K79me2 and defective STAT5 signalling in CD8⁺ T cells.

H3K79me2 is an active gene mark in mammalian cells and occurs on the promoter and 5′ regions within the coding regions of transcriptionally active genes[25,26]. Chromatin immunoprecipitation and sequencing (ChIP–seq) data[27,28] revealed high H3K79me2 occupancy in the key regulatory regions of the *STAT5B* promoter in mouse and human (Extended Data Fig. 3o, p). ChIP analysis demonstrated high levels of H3K79me2 on the *Stat5* promoter (Fig. 3p, Extended Data Table 1). This binding was diminished in T cells cultured with B16F10 supernatants and restored by methionine supplementation (Fig. 3q). Thus, H3K79me2 is involved in the direct regulation of STAT5 transcription in CD8⁺ T cells.

Methionine restores T cell immunity

To demonstrate the relevance of methionine competition between tumour cells and T cells in vivo, we conducted four complementary studies. First, we found that tumour-infiltrating CD8⁺ T cells harboured lower levels of H3K79me2 and STAT5 than did T cells in the draining lymph nodes and spleen (Extended Data Fig. 4a–d).

Second, we confirmed that human tumour-infiltrating CD8⁺ T cells from ovarian carcinoma omentum, malignant ascites, and several other cancers showed a decrease in H3K79me2 and STAT5 compared to peripheral T cells (Extended Data Fig. 4e–i). To examine the effect of methionine on human-tumour infiltrating T cells, we cultured human colorectal cancer-infiltrating T cells with or without methionine. The addition of methionine enhanced the expression of T cell effector cytokines, H3K79me2, and STAT5 (Extended Data Fig. 4j–m).

Third, we injected methionine into B16F10 tumours in mice. Methionine supplementation delayed tumour growth, enhanced H3K79me2 and STAT5 expression in tumour-infiltrating CD8⁺ T cells, and increased T cell survival and expression of polyfunctional cytokines (Fig. 4a–e). Injections of methionine into ID8 tumours in mice also slowed down

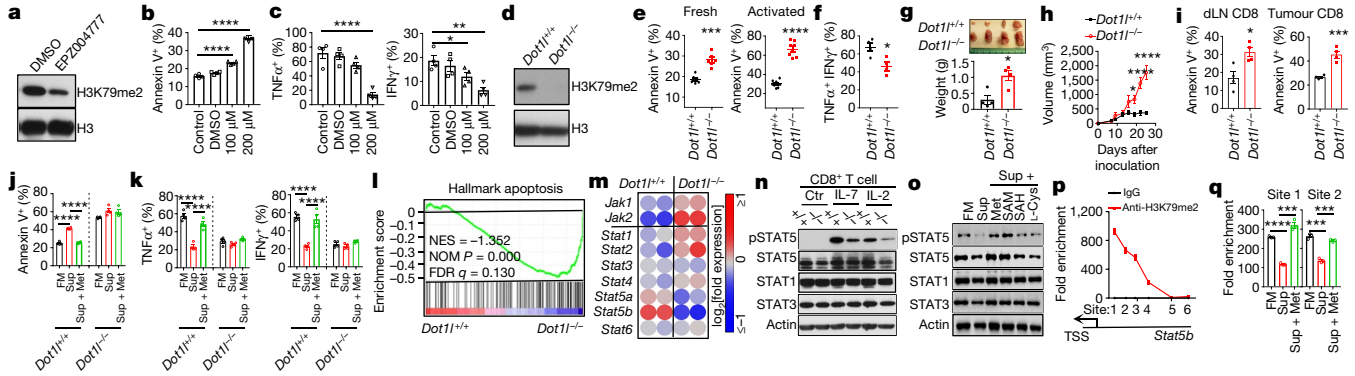


Fig. 3 | Loss of H3K79me2 impairs T cell anti-tumour immunity through STAT5. a–**c**, CD8$^+$ T cells were treated with EPZ004777 or vehicle (DMSO) for 48 h. Western blot (**a**) shows H3K79me2 in CD8$^+$ T cells. Fluorescence-activated cell sorting was used to measure apoptosis (**b**) and cytokine production (**c**). **d**, Western blot shows H3K79me2 in *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ CD8$^+$ T cells. **e**, **f**, Effect of DOT1L deletion on apoptosis (**e**) and cytokine production (**f**) in CD8$^+$ T cells. **g**–**i**, Effect of T cell DOT1L deficiency on MC38 growth (**g** (day 25), **h**) and T cell viability in draining lymph nodes (dLN) and tumour (**i**). **j**, **k**, Effect of methionine supplementation on apoptosis (**j**) and cytokine production (**k**) in *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ CD8$^+$ T cells. **l**, GSEA plot shows enriched apoptotic pathway genes in *Dot1l*$^{-/-}$ CD8$^+$ T cells. **m**, Heat map shows mRNA levels for components of the JAK–STAT pathway in mouse *Dot1l*$^{-/-}$ and *Dot1l*$^{+/+}$ CD8$^+$ T cells. **n**, Western blot shows STAT5 and p-STAT5 in *Dot1l*$^{+/+}$ (+/+) and *Dot1l*$^{-/-}$ (−/−) CD8$^+$ T cells. **o**, Western blot shows STAT5 and p-STAT5 in CD8$^+$ T cells cultured with fresh medium, tumour cell supernatant, or supernatant supplemented with different metabolites. **p**, ChIP assay shows H3K79me2 occupancy on the *Stat5b* promoter in CD8$^+$ T cells. TSS, transcription start site. **q**, ChIP assay shows H3K79me2 occupancy on the *Stat5b* promoter in CD8$^+$ T cells cultured with fresh medium, tumour cell supernatant, or supernatant supplemented with different metabolites. Data are mean ± s.e.m. Sample sizes (*n*), *P* values, statistical tests and numbers of replications are listed in 'Statistics and reproducibility' (Methods).

tumour progression (Fig. 4f), enhanced T cell (but not tumour cell) survival (Extended Data Fig. 4n), and increased effector cytokine levels in tumour ascites and tumour-infiltrating CD8$^+$ T cells (Fig. 4g, h). After methionine supplementation by injection, we detected high levels of methionine in ID8 ascites (Extended Data Fig. 4o). We also treated mice bearing CT26 tumours with methionine, anti-PD-L1, or a combination of both, and found that methionine plus anti-PD-L1 had a synergistic anti-tumour effect, compared to either treatment alone. This was accompanied with increased T cell tumour infiltration and reduced T cell apoptosis (Extended Data Fig. 4p–r).

Fourth, we provided methionine supplementation to patients with colorectal cancer (Extended Data Table 2). Methionine supplementation resulted in an increase in H3K79me2 and p-STAT5 in CD8$^+$ T cells (Fig. 4i), enhanced T cell IL-2 production (Fig. 4j) and CD8$^+$ T cell polyfunctional cytokine expression (Fig. 4k), and decreased CD8$^+$ T cell apoptosis (Fig. 4l) in these patients. Together, our data suggest that methionine deficiency impairs H3K79me2 and STAT5 expression and function in T cells.

Tumour impairs tumour immunity through SLC43A2

Methionine is transported into cells by the solute carrier family (SLC), including system L-type and A-type transporters[29]. We cultured B16F10 cells with BCH (an inhibitor of system L transporters) or MeAIB (an inhibitor of system A transporters)[29]. Then, we cultured CD8$^+$ T cells with the resulting tumour supernatants and analysed CD8$^+$ T cells. BCH, but not MeAIB, prevented T cell apoptosis and rescued the impaired cytokine profile (Extended Data Fig. 5a, b). Thus, system L transporters may be responsible for consumption of methionine by tumours.

Next, we compared the SLC transcripts in effector CD8$^+$ T cells and tumour cells, and found that *SLC7A5* and *SLC43A2* (two system L transporters) were relatively highly expressed on tumour cells (Extended Data Fig. 5c). Western blots revealed minimal SLC43A2 expression in effector CD8$^+$ T cells and comparable SLC7A5 expression in effector CD8$^+$ T cells and several tumour cells (Extended Data Fig. 5d). In line with this, we detected minimal SLC43A2 in human CD8$^+$ T cells, compared to several tumour cells (Extended Data Fig. 5e). The differential SLC43A2 expression in tumour and CD8$^+$ T cells suggests that tumour cells may be well-positioned to outcompete T cells for methionine via SLC43A2.

To test this possibility, we used a short hairpin RNA (shRNA) against *SLC43A2* (sh*SLC43A2*) to knock down *SLC43A2* in B16F10 cells (Extended Data Fig. 5f), which induced a decrease in methionine consumption (Extended Data Fig. 5g). Then, we cultured CD8$^+$ T cells with the tumour supernatants from sh*SLC43A2* cells or from tumour cells expressing scrambled shRNA. CD8$^+$ T cells cultured with supernatant from sh*SLC43A2* cells showed reduced T cell apoptosis and enhanced polyfunctional cytokine expression (Fig. 5a, b). Furthermore, T cells cultured with sh*SLC43A2* cells in a Transwell system (Extended Data Fig. 1n), showed a reduction in apoptosis and an increase in cytokine production (Fig. 5c, d), as well as increased H3K79me2, when compared with T cells cultured with control cells treated with scrambled shRNA (Fig. 5e). Thus, tumour cells outcompete T cells for methionine via SLC43A2, and this affects T cell histone methylation and function.

We next injected B16F10 cells treated with sh*SLC43A2* or scrambled shRNA (control) into *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ C57BL/6 mice. Tumour growth was slower in *Dot1l*$^{+/+}$ mice bearing sh*SLC43A2* B16F10 cells than in those bearing control B16F10 cells (Fig. 5f). However, *SLC43A2* knockdown did not affect tumour progression in *Dot1l*$^{-/-}$ mice (Extended Data Fig. 5h). The data suggest that there is a functional connection between tumour SLC43A2 and T cell DOT1L in anti-tumour immunity. Furthermore, treatment of tumour cells with sh*SLC43A2* did not affect tumour growth in *Rag1*$^{-/-}$ mice (Extended Data Fig. 5i). Thus, tumour immunity contributed to tumour control in sh*SLC43A2*-treated tumours. Consistent with this, tumour T cell infiltration (Extended Data Fig. 5j) and effector molecules were increased in sh*SLC43A2*-treated tumours when compared with control tumours in wild-type C57BL/6 mice (Fig. 5g). Treatment with anti-PD-L1 further inhibited the growth of sh*SLC43A2* B16F10 tumours (Extended Data Fig. 5k).

We also studied mice bearing sh*SLC43A2* ID8 tumours (Extended Data Fig. 5l). As with B16F10 tumours, sh*SLC43A2* had no effect on tumour growth in *Rag1*$^{-/-}$ mice (Extended Data Fig. 5m), but tumour growth was slower and CD8$^+$ T cell infiltration was increased in WT mice bearing sh*SLC43A2* ID8 tumours compared with control ID8 tumours (Extended Data Fig. 5n, o). These data suggest that pharmacologically targeting SLC43A2 may promote anti-tumour immunity. Given that no specific SLC43A2 inhibitor is available, we treated B16F10 tumour-bearing mice with BCH with or without anti-PD-L1 treatment (Fig. 5h). Treatment with BCH or anti-PD-L1 alone partially inhibited

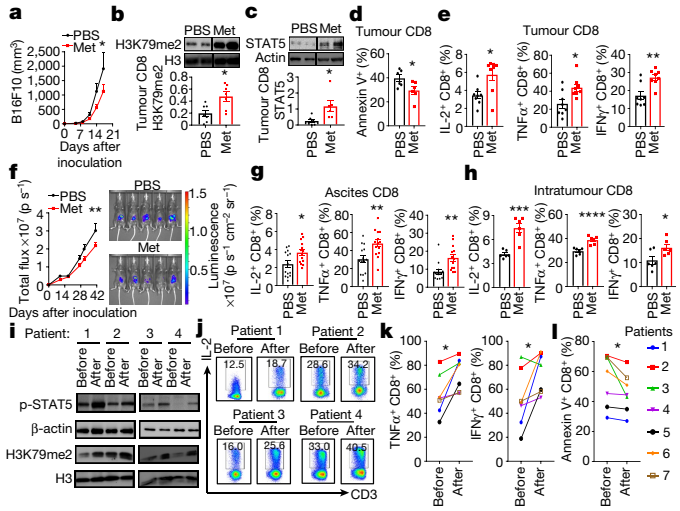


Fig. 4 | Methionine supplementation in tumours restores T cell immunity. **a**–**e**, Methionine supplementation restored T cell immunity in B16F10 tumour-bearing mice. Tumour growth (**a**), and H3K79me2 (**b**) and STAT5 (**c**) in tumour-infiltrating CD8$^+$ T cells, were monitored. FACS was used to measure apoptosis (**d**) and cytokine production (**e**) in intratumour CD8$^+$ T cells. PBS, phosphate-buffered saline (vehicle). **f**–**h**, Methionine supplementation restored T cell immunity in ID8 tumour-bearing mice. Tumour growth was monitored by bioluminescence imaging (**f**). FACS was used to measure cytokine production in CD8$^+$ T cells in the ascites (**g**) and tumours (**h**). **i**–**l**, Studies on patients with colorectal cancer treated with methionine. Western blots show p-STAT5 and H3K79me2 in peripheral CD8$^+$ T cells before and after methionine treatment (**i**). FACS was used to measure IL-2$^+$ T cells (**j**) and effector cytokine production (**k**) and apoptosis (**l**) in CD8$^+$ T cells in patients before and after methionine treatment. Data are mean ± s.e.m. Sample sizes (*n*), *P* values, statistical tests and numbers of replications are listed in 'Statistics and reproducibility' (Methods).

tumour growth, and the combination had a synergistic effect (Fig. 5h). Moreover, the combination treatment induced the highest T cell infiltration (Extended Data Fig. 5p) and the highest expression of effector molecules in tumour-infiltrating CD8$^+$ T cells (Fig. 5i). The combination treatment also synergistically inhibited tumour growth and enhanced cytokine production by tumour-infiltrating CD8$^+$ T cells in mice bearing ID8 tumours (Fig. 5j–l). These results suggest that targeting tumour SLC43A2 in combination with checkpoint blockade may be an effective anti-cancer approach.

Finally, we investigated a potential relationship between tumour SLC43A2 expression, T cell signatures, and clinical outcome in patients with cancer. Using data from the Cancer Genome Atlas (TCGA) database, we found that *SLC43A2* transcripts were higher in tumours than in matched normal tissue (Fig. 5m). Moreover, high tumour SLC43A2 expression was associated with poor survival (Extended Data Fig. 5q–s). Single-cell RNA-seq of tumour-infiltrating T cells and cancer cells from patients with melanoma[30] showed higher *SLC43A2* transcripts in tumour cells than in T cells (Extended Data Fig. 5t). We divided the patients into two groups according to high or low expression of SLC43A2 in tumours. GSEA showed that expression of methionine metabolic signalling genes was enriched in melanoma cells with high levels of SLC43A2 (Extended Data Fig. 5u). Tumour SLC43A2 levels negatively correlated with *CD8* and *IFNG* transcripts in the same tumours (Extended Data Fig. 5v). Furthermore, compared to patients with low tumour SLC43A2 levels, T cells in patients with high tumour SLC43A2 levels showed weak signatures for methionine metabolism and histone methylation, and low expression of effector genes (Extended Data Fig. 5w–y). Thus, high expression of SLC43A2 in tumours is associated with reduced T cell immune responses in patients with cancer.

Discussion

Recent studies have started to explore the roles of amino acids in T cell activation and epigenetic reprogramming[31–33]. Methionine is an essential amino acid, and is converted to SAM for methyltransferases to yield methylated substrates, including histone methylation[18,19]. Hence, SAM provides a link between methionine metabolism and epigenetic regulation. Dysfunctional T cells exhibit a distinct epigenetic landscape including histone alteration[2,3]. Thus, abnormal methionine metabolism may lead to particular histone alteration in T cells and contribute to their dysfunction in the tumour microenvironment.

We have demonstrated direct competition between tumour cells and T cells for methionine, which results in reductions in a series of substrates for one-carbon metabolism, including SAM, in T cells. DOT1L is the only methyltransferase for H3K79[20,21], and has a relatively low Michaelis constant (K_m) for SAM[34]. These characteristics contribute

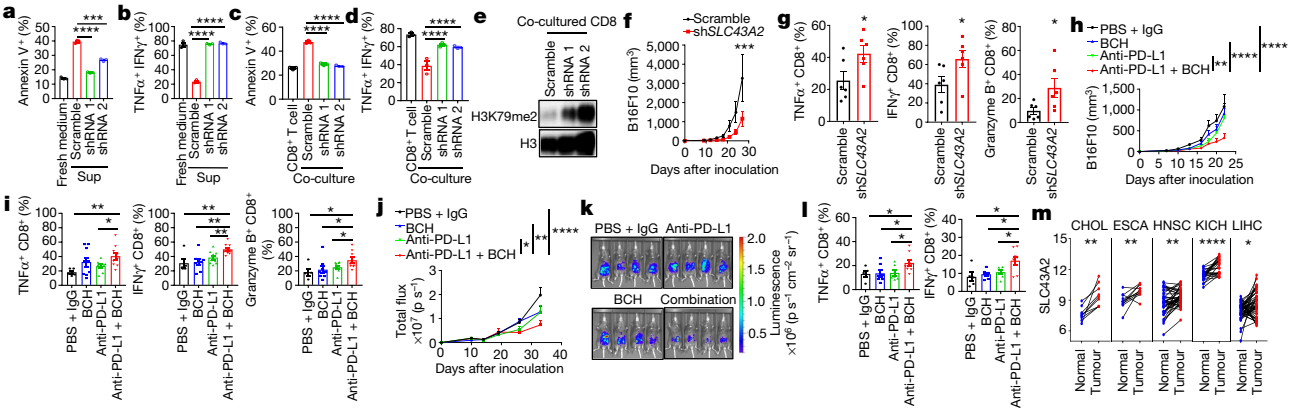


Fig. 5 | Tumour cells outcompete T cells for methionine via SLC43A2. a, **b**, Effects of supernatants from sh*SLC43A2*-treated tumours on T cell apoptosis (**a**) and cytokine production (**b**). Scramble, control; shRNA 1 and shRNA 2, sh*SLC43A2*. **c**–**e**, Effects of tumour *SLC43A2* knockdown on apoptosis (**c**), cytokine production (**d**) and H3K29me2 modification (**e**) in T cells. CD8$^+$ T cells were co-cultured with wild-type and sh*SLC43A2*-treated B16F10 cells in a Transwell system in medium containing 30 μM methionine. Apoptosis (**c**), cytokine production (**d**) and H3K79me2 (**e**) in CD8$^+$ T cells were determined by FACS and western blot after 72 h. **f**, **g**, Effects of tumour *SLC43A2* knockdown on tumour growth (**f**) and tumour T cell function (**g**). **h**–**l**, Effects of a combination of BCH and anti-PD-L1 treatment on B16F10 (**h**, **i**) and ID8 (**j**–**l**) tumour-bearing mice. **h**, **j**, **k**, Tumour growth; **i**, **l**, TNFα$^+$, IFNγ$^+$ and granzyme B$^+$ CD8$^+$ T cells. **m**, *SLC43A2* transcripts in tumours and paired adjacent normal tissue samples for several types of tumour from TCGA. CHOL, cholangiocarcinoma; ESCA, oesophageal carcinoma; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; LIHC, liver hepatocellular carcinoma. Data are mean ± s.e.m. Sample sizes (*n*), *P* values, statistical tests and numbers of replications are listed in 'Statistics and reproducibility' (Methods).

to the sensitivity of H3K79 methylation to deprivation of methionine and SAM, and explain why T cell H3K79me2 is predominantly sensitive to tumour-altered methionine metabolism. We have also validated the mechanistic connection between methionine, H3K79me2, and STAT5 in human and mouse tumour-infiltrating T cells. Notably, as substantial methionine is required for abnormal tumour cell proliferation and metabolism[35], the effects of dietary methionine restriction on tumour growth have been tested in immune-deficient systems[36]. Our work indicates that both human and mouse effector T cells are sensitive to methionine. Thus, tumour-specific methionine restriction is essential to maintain T cell immunity in patients with cancer.

H3K79me2 is an active transcriptional histone mark[21]. Biochemical and genetic inhibition of DOT1L abolished H3K79me2 and STAT5, resulting in T cell apoptosis and dysfunction. Mechanistically, H3K79me2 controls STAT5 transcription in $CD8^+$ T cells. Thus, we have identified a causal and biological link between a particular histone alteration (H3K79me2) and a transcription factor (STAT5) that is critical for defining T cell phenotype. Moreover, tumour cells outcompete T cells for methionine via SLC43A2, a major methionine transporter. Given that SLC43A2 is highly expressed on multiple human and mouse tumour cells with different genetic backgrounds, abnormal tumour SLC43A2 and its related methionine metabolism are unlikely to be driven by shared key oncogenes. Inhibition of tumour SLC43A2 can normalize methionine metabolism in effector T cells and rescue their function, and can also improve spontaneous and checkpoint blockade-induced anti-tumour immunity in preclinical models. Our work has not only generated insights into SLC biology in T cells, but also identified tumour SLC43A2 as an mechanism of immunotherapy resistance in patients with cancer. Indeed, we have established a negative correlation between tumour cell SLC43A2 levels, histone methylation in tumour-infiltrating T cells, and effector functional signatures in the same tumour tissues in patients with cancer.

In summary, we have shown that tumour cells metabolically and epigenetically impair T cell function and tumour immunity by outcompeting T cells for methionine via SLC43A2 (Extended Data Fig. 6). Our work demonstrates a long-suspected crosstalk between metabolism, histone pattern, and functional profile in tumour-infiltrating T cells. On this basis, selectively targeting tumour methionine metabolism may be a useful approach for cancer immunotherapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2682-1.

1. Wherry, E. J. T cell exhaustion. *Nat. Immunol.* **12**, 492–499 (2011).
2. Pauken, K. E. et al. Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science* **354**, 1160–1165 (2016).
3. Sen, D. R. et al. The epigenetic landscape of T cell exhaustion. *Science* **354**, 1165–1169 (2016).
4. Philip, M. et al. Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* **545**, 452–456 (2017).
5. Zou, W., Wolchok, J. D. & Chen, L. PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: mechanisms, response biomarkers, and combinations. *Sci. Transl. Med.* **8**, 328rv4 (2016).
6. Horton, B. L., Williams, J. B., Cabanov, A., Spranger, S. & Gajewski, T. F. Intratumoral $CD8^+$ T-cell apoptosis is a major component of T-cell dysfunction and impedes antitumor immunity. *Cancer Immunol. Res.* **6**, 14–24 (2018).
7. Schoenborn, J. R. et al. Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-γ. *Nat. Immunol.* **8**, 732–742 (2007).
8. Khan, O. et al. TOX transcriptionally and epigenetically programs $CD8^+$ T cell exhaustion. *Nature* **571**, 211–218 (2019).
9. Zhao, E. et al. Cancer mediates effector T cell dysfunction by targeting microRNAs and EZH2 via glycolysis restriction. *Nat. Immunol.* **17**, 95–103 (2016).
10. Song, M. et al. IRE1α-XBP1 controls T cell function in ovarian cancer by regulating mitochondrial activity. *Nature* **562**, 423–428 (2018).
11. Li, W. et al. Aerobic glycolysis controls myeloid-derived suppressor cells and tumor immunity via a specific CEBPB isoform in triple-negative breast cancer. *Cell Metab.* **28**, 87–103.e6 (2018).
12. Maj, T. et al. Oxidative stress controls regulatory T cell apoptosis and suppressor activity and PD-L1-blockade resistance in tumor. *Nat. Immunol.* **18**, 1332–1341 (2017).
13. Levine, A. J. & Puzio-Kuter, A. M. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* **330**, 1340–1344 (2010).
14. Cairns, R. A., Harris, I. S. & Mak, T. W. Regulation of cancer cell metabolism. *Nat. Rev. Cancer* **11**, 85–95 (2011).
15. Guttormsen, A. B., Solheim, E. & Refsum, H. Variation in plasma cystathionine and its relation to changes in plasma concentrations of homocysteine and methionine in healthy subjects during a 24-h observation period. *Am. J. Clin. Nutr.* **79**, 76–79 (2004).
16. Schmidt, J. A. et al. Plasma concentrations and intakes of amino acids in male meat-eaters, fish-eaters, vegetarians and vegans: a cross-sectional analysis in the EPIC-Oxford cohort. Eur. *J. Clin. Nutr.* **70**, 306–312 (2016).
17. Qiu, J. et al. Acetate promotes T cell effector function during glucose restriction. *Cell Rep.* **27**, 2063–2074.e5 (2019).
18. Mentch, S. J. et al. Histone methylation dynamics and gene regulation occur through the sensing of one-carbon metabolism. *Cell Metab.* **22**, 861–873 (2015).
19. Shiraki, N. et al. Methionine metabolism regulates maintenance and differentiation of human pluripotent stem cells. *Cell Metab.* **19**, 780–794 (2014).
20. Min, J., Feng, Q., Li, Z., Zhang, Y. & Xu, R.-M. Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase. *Cell* **112**, 711–723 (2003).
21. Nguyen, A. T. & Zhang, Y. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev.* **25**, 1345–1358 (2011).
22. Jo, S. Y., Granowicz, E. M., Maillard, I., Thomas, D. & Hess, J. L. Requirement for Dot1l in murine postnatal hematopoiesis and leukemogenesis by MLL translocation. *Blood* **117**, 4759–4768 (2011).
23. Villarino, A., Kanno, Y. & O'Shea, J. Mechanisms and consequences of Jak–STAT signaling in the immune system. *Nat. Immunol.* **18**, 374–384 (2017).
24. Kagoya, Y. et al. DOT1L inhibition attenuates graft-versus-host disease by allogeneic T cells in adoptive immunotherapy models. *Nat. Commun.* **9**, 1915 (2018).
25. Schübeler, D. et al. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
26. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
27. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
28. Rosenbloom, K. R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
29. Alexander, S. P. H. et al. The concise guide to pharmacology 2019/20: transporters. *Br. J. Pharmacol.* **176** (Suppl. 1), S397–S493 (2019).
30. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
31. Ma, E. H. et al. Serine is an essential metabolite for effector T cell expansion. *Cell Metab.* **25**, 345–357 (2017).
32. Geiger, R. et al. l-Arginine modulates T cell metabolism and enhances survival and anti-tumor activity. *Cell* **167**, 829–842.e13 (2016).
33. Roy, D. G. et al. Methionine metabolism shapes T helper cell responses through regulation of epigenetic reprogramming. *Cell Metab.* **31**, 250–266.e9 (2020).
34. Richon, V. M. et al. Chemogenetic analysis of human protein methyltransferases. *Chem. Biol. Drug Des.* **78**, 199–210 (2011).
35. Cavuoto, P. & Fenech, M. F. A review of methionine dependency and the role of methionine restriction in cancer growth control and life-span extension. *Cancer Treat. Rev.* **38**, 726–736 (2012).
36. Gao, X. et al. Dietary methionine influences therapy in mouse cancer models and alters human metabolism. *Nature* **572**, 397–401 (2019).

Methods

Mice

Six- to eight-week-old female wild-type C57BL/6, BALB/c, and *Rag1* knockout (KO) mice were obtained from the Jackson Laboratory. *Dot1l*$^{flox/flox}$ mice were bred with CD4-Cre mice to generate mice with specific *Dot1l* deletion in T cells. All *Dot1l*$^{+/+}$ or *Dot1l*$^{-/-}$ mice were used at the age of 6–12 weeks unless specified in the text. Mice were housed under specific pathogen-free conditions and handled according to the guidelines of the University Committee on the Use and Care of Animals at the University of Michigan.

Clinical studies

Patients with colorectal cancer were recruited for the methionine supplementation study. Eligible patients were of Eastern Cooperative Oncology Group performance status 0/1 with adequate organ and bone marrow function. Patients were excluded from this study if they had received or were receiving any concurrent chemotherapy, immunotherapy, biologic, and/or hormonal therapy for cancer. All patients took two capsules of methionine (500 mg per capsule, NOW Foods) daily for two weeks. This study was conducted according to the Declaration of Helsinki and approved by the institutional review board (IRB) of the Medical University of Lublin, with written informed consent obtained from all patients. Study participants were not compensated.

Human specimens

Plasma, peripheral blood mononuclear cells (PBMCs), and tumour-infiltrating T cells were isolated from healthy donors and patients with cancer. Plasma from patients diagnosed with high-grade serous ovarian carcinomas were collected for this study. Human specimens were collected with informed consent and procedures approved by the IRB of the University of Michigan.

Reagents

Amino acids, including L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-threonine, L-tryptophan, L-valine, L-histidine, L-arginine, L-cystine, L-tyrosine, and MEM non-essential amino acid solution (100×, including L-alanine, L-aspartic acid, L-asparagine, L-glutamic acid, L-glycine, L-proline and L-serine) were from Sigma. L-Glutamine (100×), 2-mercaptoethanol and dialysed fetal bovine serum (FBS) were from GIBCO. RPMI 1640 medium without amino acids and sodium phosphate powder (#R8999-04A) were from US Biological. 1× RPMI without L-glutamine, L-cysteine, L-cystine, and L-methionine was from MP Biomedicals. Methionine assay kit (Fluorometric) was from Abcam (ab234041). Anti-mouse CD3 and anti-CD28, anti-human CD3 and anti-CD28 monoclonal antibodies (mAbs) were from eBioscience. Mouse and human interleukin 2 (IL-2) were from R&D Systems, Inc.. *S*-(5-adenosyl)- L-methionine iodide (SAM), *S*-(5-adenosyl)- L-homocysteine (SAH), and L-cystathionine, as well as SLC transporter inhibitors, including α-(methylamino) isobutyric acid (MeAIB) and 2-amino-2-norbornanecarboxylic acid (BCH) were from Sigma. The DOT1L inhibitor EPZ004777 (#1338466-77-5) was from Millipore. Anti-mouse PD-L1 (clone: 10F.9G2) and rat IgG2B isotype (clone: LTF-2) were from Bioxcell.

Cell separation and culture

Human cells (including A375, CHL-1, SK-MEL-2 and 293T cells) and mouse tumour cells (including B16F10 and CT26 cells) were obtained from the ATCC. Mouse ID8-luc and MC38 cells were as previously reported[37,38]. Human primary high-grade serous ovarian carcinoma cells (OC8) were generated in our laboratory[38]. All cell lines in our laboratory are routinely tested for mycoplasma contamination and cells used in this study were negative for mycoplasma. None of our cell lines are on the list of commonly misidentified cell lines (International Cell Line Authentication Committee). Tumour cells were maintained in RPMI1640 (HyClone SH30255, GE Healthcare, Chicago, IL, USA) containing 10% (v/v) FBS (Alkali Scientific) and 1% (v/v) pen/strep (GIBCO).

Mouse lymphocytes were isolated from spleen and lymph nodes. CD8$^+$ T cells were separated using the EasySep Mouse CD8$^+$ T Cell Isolation Kit (STEMCELL Technologies Inc.). Human PBMCs were isolated from blood using Lymphoprep (STEMCELL Technologies Inc.). Human CD8$^+$ T cells were separated using the EasySep Human CD8$^+$ T Cell Isolation Kit (STEMCELL Technologies Inc.). CD8$^+$ T cells were re-suspended (10^6 cells/ml) and activated with anti-CD3 and anti-CD28 mAbs for 48 h. Activated CD8$^+$ T cells were maintained with IL-2 (10 ng/ml) and 2-mercaptoethanol, and cultured with fresh complete medium, media with or without individual amino acids, or tumour supernatants for 36–48 h. Media with or without amino acids were formulated with RPMI1640 (US Biological, #R8999-04A) by supplementation or omission of individual amino acids. The tumour supernatants were collected from medium initially cultured with tumour cells in RPMI lacking L-glutamine, L-cysteine, L-cystine, and L-methionine (MP Biomedicals #1646454), and subsequently supplemented with L-glutamine (GIBCO), L-cystine 2HCl (Sigma), and different concentrations of L-methionine (Sigma), as specified in different experiments. The following amino acids were used to supplement tumour culture supernatants: Iso 380 μM/l, Leu 380 μM/l, Lys 220 μM/l, Met 30 μM/l, Phe 90 μM/l, Thr 170 μM/l, Trp 20 μM/l, Val 170 μM/l, His 100 μM/l, Arg 1,160 μM/l, Cys 200 μM/l, Gln 2,000 μM/l, Tyr 110 μM/l, and other amino acids (MEM Non-essential Amino Acid Solution 100×, a mix of Ala, Asp, Asn, Glu, Gly, Pro, and Ser (Sigma)). The following metabolites were used to supplement tumour culture supernatants: Met 30 μM/l, SAM 50 μM/l, SAH 50 μM/l or L-cystathionine 100 μM/l.

Intratumour CD8$^+$ T cells from mice and humans were isolated as follows: mononuclear cells from the whole tumour or ascites suspension were first enriched by density gradient centrifugation using Lymphoprep (STEMCELL). CD8$^+$ T cells were further separated by a negative–positive two-step isolation. First, the enriched cells were isolated using the EasySep Mouse/Human CD8$^+$ T Cell Isolation Kit (negative selection, STEMCELL), then further enriched using the EasySep Mouse/Human CD8 Positive Selection Kit II (positive selection, STEMCELL). The purity of CD8$^+$ T cells was further determined by FACS staining.

Generation of knockdown cells

SLC43A2 knockdown cells were generated using the MISSION shRNA (Sigma) and GIPZ Lentiviral shRNA systems (Dharmacon, Inc.). 293T cells were co-transfected with lentiviral shRNA system together with plasmids psPAX2 and pMD2.G using Lipofectamine 2000 (Thermo Fisher Scientific) for lentivirus package. Forty-eight hours after transfection, the supernatant was collected. Tumour cells were infected with the virus supernatant for 24 h and then selected with 2 μg/ml puromycin (Santa Cruz Biotechnology) for an additional 48 h. Knockdown efficiency was validated by immunoblotting.

Flow cytometry analysis (FACS)

For FACS staining, cells were stained with a combination of fluorescence-conjugated mAbs from BD Biosciences or Thermo Fisher Scientific (Waltham). Mouse samples were stained with FITC-Annexin V, 7-AAD, PE-texas red-anti mouse CD45 (30-F11), FITC-anti-mouse CD90 (53-2.1), APC-Cy7-anti-mouse CD4 (RM4-5), AF700-anti-mouse CD8 (53-6.7), APC-anti-mouse IL-2 (JES6-5H4), BV786-anti-mouse IFNγ (XMG1.2), PE-Cy7-anti-mouse TNFα (MP6-XT22) and PE-anti-mouse granzyme B (NGZB) mAbs. Human samples were stained with FITC-Annexin V, 7-AAD, pacific blue-anti-human IFNγ (4S.B3), and APC-anti-human TNFα (MAb11). For apoptosis staining, cells were washed with 1× binding buffer (BD Biosciences) and stained with Annexin V and 7-AAD in 1× binding buffer in the dark for 10 min. For surface staining, the cells were incubated in the dark with antibodies for 30 min. For intracellular staining, the cells were fixed in Fix/Perm solution (BD Biosciences). After being washed with Perm/Wash buffer

(BD Biosciences), the cells were stained intracellularly for 30 min in the dark. For STAT5, cells were stained with APC-anti-STAT5 (REA549, Miltenyi Biotec Inc.). For DOT1L and H3K79me2 intracellular staining, the cells were first stained with DOT1L or H3K79me2 antibodies (Abcam), and then stained using a FITC-conjugated goat anti-rabbit IgG (H + L) secondary antibody (Invitrogen). All samples were acquired on BD LSRFortessa (BD Biosciences) and were analysed using FACSDiva (BD Biosciences) or FlowJo (FlowJo LLC).

Chromatin immunoprecipitation

ChIP assay was performed according to the manufacturer's protocol (Millipore). In brief, crosslinking was performed with 1% formaldehyde or 1% paraformaldehyde for 10 min. To enhance cell lysis, we ran the lysate through a 27-g needle three times and flash froze it at −80 °C. Sonication was then performed using the the Misonix 4000 water bath sonication unit at 15% amplitude for 20 min. Protein–DNA complex was precipitated by specific antibodies against H3K79me2 (Abcam) and IgG control (Millipore). DNA was purified using DNA Purification Kit (Qiagen). ChIP-enriched chromatin was used for real-time PCR. Relative expression levels were normalized to Input. Specific primers are listed in Extended Data Table 1.

Real-time PCR and western blotting

$CD8^+$ T cells were incubated in fresh medium or specific tumour supernatant for 24–48 h. Tumour cells were maintained with complete medium. The cells were washed and collected. RNA was isolated from these cells using Direct-zolTM RNA miniprep Plus kit (ZOMO research), and then subjected to reverse transcription using first-strand cDNA Synthesis for Quantitative RT–PCR kit (OriGene). Real-time PCR was performed using SYBR green chemistry (Applied Biosystems). Reactions were run on a real-time PCR system (StepOnePlus Real-Time PCR System, Applied Biosystem). Specific primers are listed in Extended Data Table 3.

For western blotting, $CD8^+$ T and tumour cells were washed and lysed in a modified RIPA buffer with 1× protease inhibitor cocktail (Roche). The lysates were stored at −80 °C until immunoblot analysis. For histone isolation, $CD8^+$ T cells were first lysed with PBS with 0.5% Triton X-100. The lysates were then incubated on ice for 10 min and cleared by centrifugation at 5,000*g* for 15 min. The precipitate was dissolved with 0.2 N HCl. Protein concentration was quantified using a BCA protein assay kit (Thermo Fisher Scientific) and denatured at 95 °C for 5 min. The lysate samples were stored at −80 °C for immunoblot analysis. In brief, the proteins were separated electrophoretically using a 12% SDS-polyacrylamide gel and transferred onto an PVDF membrane (Millipore). The membranes were blocked in 5% fat-free milk for 1 h, and then incubated with a specific primary antibody at 4 °C overnight. Blots were probed with rabbit anti-H3K79me2, H3K4me2, H3K4me3, H3K9me2, H3K27me2, total H3, STAT5, p-STAT5, STAT1, STAT3, SLC43A2, SLC7A5 and β-actin antibodies. All antibodies were from Abcam or Cell Signaling Technology. After hybridization with HRP-conjugated secondary antibody (Life Technologies), protein bands were visualized using a chemiluminescence detection kit (Bio-Rad Laboratories).

Tumour inoculation and treatments

For the in vivo tumour growth experiments, mice were inoculated subcutaneously with 2×10^5 B16F10 cells or 5×10^5 MC38 cells, or intraperitoneally with 2×10^6 ID8-Luc ovarian cancer cells. The B16F10 and MC38 tumour volumes were measured along three orthogonal axes (*a*, *b*, and *c*) and were calculated as follows: tumour volume = $a \times b \times c/2$. The ID8-luc tumour growth was monitored using the Xenogen IVIS Spectrum In vivo Bioluminescence Imaging System (PerkinElmer). Tumour load was calculated based on the total flux (photons per second). Anti-PD-L1 and IgG1 isotype mAbs (Bioxcell) were given intraperitoneally at a dose of 100 μg per mouse on day 7 after tumour cell inoculation, then every 3 days for the duration of the experiment. 2-Amino-2-norbornanecarboxylic acid (BCH) was given intravenously at a dose of 180 mg/kg per mouse on day 7 after tumour inoculation, then every 2 days for the duration of the experiment. Methionine was given by intratumour (B16F10 model) or intraperitoneal (ID8 model) injection at a dose of 40 mg/kg per mouse on day 7 after tumour inoculation, then every 2 days for the duration of the experiment. Animal studies were conducted under the approval of the University of Michigan Committee on Use and Care of Animals. In none of the experiments did tumour size surpass 2 cm in any dimension. No mouse had severe abdominal distension (≥10% original body weight increase). Sample size was chosen on the basis of preliminary data. After tumour inoculation, mice were randomized and assigned to different groups for treatment.

RNA-seq and bioinformatics analysis

$CD8^+$ T cells were cultured in complete fresh medium, tumour supernatant (sup), or tumour supernatant plus methionine (sup + met) for 24 h. $CD8^+$ T cells from *Dot1l*$^{-/-}$ mice and littermates (*Dot1l*$^{+/+}$) were isolated and sorted. The RNA was isolated by using Direct-zolTM RNA miniprep Plus kit (ZOMO research). RNA-seq and RNA-array were conducted in $CD8^+$ T cells by the DNA sequencing core at the University of Michigan. The data were processed by the Bioinformatic Core at the University of Michigan, and analysed with ClueGo[39] and GSEA software v. 3.0[40]. RNA sequencing data that support the findings of this study have been deposited in NCBI Gene Expression Omnibus (GEO) under accession number GSE150887. Public RNA-seq data were from GSE108694 and GSE72056. For single cell RNA-seq data in patients with melanoma, the expression levels of genes were quantified as $E_{i,j} = \log_2(\mathrm{TPM}_{i,j}/10 + 1)$, where $\mathrm{TPM}_{i,j}$ refers to transcripts per million (TPM) for gene *i* in sample *j*[30]. We then evaluated the average $E_{i,j}$ values of tumour cell *SLC43A2* transcripts[41–43]. Based on the median of average $E_{i,j}$ values, we divided the patients into high (average SLC43A2 $E_{i,j} > 0.056$, including patients 53, 79, 81, 82, 84, and 94) and low (average SLC43A2 $E_{i,j} < 0.056$, including patients 60, 65, 71, 80, 88, and 89) groups. GSEA analysis for tumour-infiltrating T cells was characterized and compared between patients with high and low levels of *SLC43A2* expression.

Metabolomics

Metabolomics and sample collection were performed as previously reported[44,45]. In brief, $CD8^+$ T cells were collected and transferred to a 15-ml tube for centrifugation at 300*g* for 5 min at 4 °C. Cells were then washed with cold PBS. After further centrifugation at 300*g* for 5 min at 4 °C, 1 ml of 80% cold methanol was added and vigorously vortexed to ensure the cell pellet was completely disrupted. The samples were placed on dry ice and moved to a −80 °C freezer for 10 min, followed by vigorous vortexing. The samples were again centrifuged at maximum speed for 10 min at 4 °C. The supernatant was collected in new tubes and normalized by protein concentration. Samples were kept at −80 °C until analysis. We used liquid chromatography–mass spectrometry to detect intracellular metabolites in $CD8^+$ T cells and amino acids in human serum from healthy donors and patients with ovarian cancer. Intracellular SAM and amino acids in B16F10 supernatants were measured in Creative-proteomics.

Statistical analysis

Statistical analysis was performed using GraphPad Prism statistical software (version 7, GraphPad Software Inc.). Error bars in data represent s.e.m. Inter-group data were analysed using an unpaired or paired two-tailed *t*-test. Tumour growth was analysed using two-way ANOVA. Survival functions were estimated using the Kaplan–Meier method. A log-rank test was used to calculate the statistical differences. The correlations between tumour *SLC43A2* and immune-associated genes were analysed using Pearson's correlation test. A value of $P < 0.05$ was considered statistically significant. Experiments were not randomized unless otherwise stated, and experimenters were not blinded to allocation during experiments and outcome assessment.

Statistics and reproducibility

Figure 1a–d, $n = 3$ biologically independent samples. The experiments were performed three times with similar results. a, $^{****}P < 0.0001$ compared with complete medium (CM) by two-tailed unpaired *t*-test; b, $^{***}P = 0.001$ compared with CM by two-tailed unpaired *t*-test; c, $^{****}P < 0.0001$ compared with CM by two-tailed unpaired *t*-test, d, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. e, $n = 4$ biologically independent samples. The experiments were performed three times with similar results. $^{**}P = 0.0014$, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. f, $n = 3$ biologically independent samples. The experiments were performed once. g–j, $n = 3$ biologically independent samples. The experiments were performed three times with similar results. g, $^{****}P < 0.0001$ compared with supernatant (sup) by two-tailed unpaired *t*-test. h, $^{***}P = 0.0002$, $^{****}P < 0.0001$ compared with sup by two-tailed unpaired *t*-test. i, NS $P = 0.6597$, $^{***}P = 0.0007$, and $^{***}P = 0.0002$ by two-tailed unpaired *t*-test. k, l, $n = 4$ biologically independent samples. The experiments were performed three times with similar results.

Figure 2a, b, $n = 4$ biologically independent samples. $CD8^+$ T cells were cultured with fresh medium (FM), B16F10 tumour supernatants (sup) and B16F10 supernatants with methionine supplementation (sup + met). GSEA plot showed enriched apoptotic and defective TCR signaling pathways (a), as well as defective methionine metabolism signalling (b) in T cells cultured in tumour supernatant. c–f, $n = 4$ biologically independent samples. The experiments were performed three times with similar results. d, $^{***}P = 0.0006$ FM vs sup, and $^{***}P = 0.0005$ sup vs sup + met by two-tailed unpaired *t*-test. e, $^{****}P < 0.0001$ FM vs sup, and $^{*}P = 0.0183$ sup vs sup + met by two-tailed unpaired *t*-test. f, $^{****}P < 0.0001$ FM vs sup, and $^{**}P = 0.0082$ sup vs sup + met by two-tailed unpaired *t*-test. g, h, $n = 3$ biologically independent samples. The experiments were performed three times with similar results. $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. i, j, The experiments were performed three times with similar results.

Figure 3a, The experiments were performed three times with similar results. b, c, $n = 4$ biologically independent samples. The experiments were performed three times with similar results. b, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. c, $^{****}P < 0.0001$, $^{*}P = 0.0497$, $^{**}P = 0.0025$ by two-tailed unpaired *t*-test. d, The experiments were performed three times with similar results. e, $n = 6$ biologically independent samples in fresh $CD8^+$, and $n = 8$ biologically independent samples in activated $CD8^+$. The experiments were performed twice with similar results. $^{***}P = 0.0001$, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. f, $n = 4$ biologically independent samples. The experiments were performed three times with similar results. $^{*}P = 0.0165$ by two-tailed unpaired *t*-test. g–i, $n = 4$ biologically independent samples. The experiments were performed twice with similar results. g, $^{*}P = 0.0109$ by two-tailed unpaired *t*-test. h, Day 19 $^{*}P = 0.0169$, Day 22 and Day 25 $^{****}P < 0.0001$ by two-way ANOVA. i, $^{*}P = 0.014$, $^{***}P = 0.0008$ by two-tailed unpaired *t*-test. j, k, $n = 4$ biologically independent samples. The experiments were performed twice with similar results. $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. l, m, $n = 3$ biologically independent samples. The data were from RNA array of *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ $CD8^+$ T cells. n, o, The experiments were performed three times with similar results. p, q, $n = 3$ biologically independent samples. The experiments were performed three times with similar results. q, Site1: $^{***}P = 0.0002$ sup vs sup + met, $^{****}P < 0.0001$ sup vs FM by two-tailed unpaired *t*-test. Site2: $^{***}P = 0.0002$ sup vs FM, $^{***}P = 0.0001$ sup vs sup + met by two-tailed unpaired *t*-test.

Figure 4a, $n = 10$ biologically independent animals. The experiments were performed once. Day 17 $^{*}P = 0.0291$ by two-tailed unpaired two-way ANOVA. b, $n = 7$ biologically independent animals. $^{*}P = 0.0114$ by two-tailed unpaired *t*-test. c, $n = 6$ biologically independent animals. $^{*}P = 0.0292$ by two-tailed unpaired *t*-test. d, $n = 6$ biologically independent animals. $^{*}P = 0.0395$ by two-tailed unpaired *t*-test. e, $n = 8$ biologically independent animals. IL-2: $^{*}P = 0.0106$ by two-tailed unpaired *t*-test. TNFα: $^{*}P = 0.0131$ by two-tailed unpaired *t*-test. IFNγ: $^{**}P = 0.0027$ by two-tailed unpaired *t*-test. f, $n = 20$ biologically independent animals. The experiments were performed once. Day 41 $^{*}P = 0.0046$ by two-tailed unpaired two-way ANOVA. g, $n = 14$ biologically independent animals. IL-2: $^{*}P = 0.0161$ by two-tailed unpaired *t*-test. TNFα: $^{**}P = 0.0046$ by two-tailed unpaired *t*-test. IFNγ: $^{**}P = 0.0024$ by two-tailed unpaired *t*-test. h, $n = 6$ biologically independent samples. IL-2: $^{***}P = 0.0004$ by two-tailed unpaired *t*-test. TNFα: $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. IFNγ: $^{*}P = 0.0328$ by two-tailed unpaired *t*-test. i–l, $n = 7$ patients with colorectal cancer received methionine supplementation. i, j, Representative results were shown from four patients. k, l, TNFα: $^{*}P = 0.0181$ by two-tailed paired *t*-test. IFNγ: $^{*}P = 0.0404$ by two-tailed paired *t*-test. Annexin V: $^{*}P = 0.0497$ by two-tailed paired *t*-test.

Figure 5a–d, $n = 3$ biologically independent animals. The experiments were performed three times with similar results. a, $^{***}P = 0.0001$, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. b-d, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. e, The experiments were performed three times with similar results. f, $n = 8$ biologically independent animals. The experiments were performed twice with similar results. Day 27 $^{***}P = 0.0006$ by two-tailed unpaired two-way ANOVA. g, $n = 6$ biologically independent animals. The experiments were performed twice with similar results. TNFα: $^{*}P = 0.0494$ by two-tailed unpaired *t*-test. IFNγ: $^{**}P = 0.0496$ by two-tailed unpaired *t*-test. Granzyme B: $^{*}P = 0.0434$ by two-tailed unpaired *t*-test. h, $n = 18$ biologically independent animals. The experiments were performed twice with similar results. Day 22 $^{**}P = 0.0023$ anti-PD-L1 + BCH vs anti-PD-L1, $^{****}P < 0.0001$ anti-PD-L1 + BCH vs PBS + IgG or BCH by two-tailed unpaired two-way ANOVA. i, $n = 5$ in PBS + IgG, $n = 9$ in other groups biologically independent animals. The experiments were performed twice with similar results. TNFα: $^{*}P = 0.0263$ anti-PD-L1 + BCH vs anti-PD-L1, $^{**}P = 0.0082$ anti-PD-L1 + BCH vs PBS + IgG by two-tailed unpaired *t*-test. IFNγ: $^{**}P = 0.0046$ anti-PD-L1 + BCH vs PBS + IgG, $^{**}P = 0.0013$ anti-PD-L1 + BCH vs BCH, $^{**}P = 0.0086$ anti-PD-L1 + BCH vs anti-PD-L1 by two-tailed unpaired *t*-test. Granzyme B: $^{*}P = 0.0235$ anti-PD-L1 + BCH vs PBS + IgG, $^{**}P = 0.0426$ anti-PD-L1 + BCH vs BCH, $^{**}P = 0.0484$ anti-PD-L1 + BCH vs anti-PD-L1 by two-tailed unpaired *t*-test. j, k, $n = 9$ biologically independent animals. The experiments were performed twice with similar results. Day 33 $^{*}P = 0.0167$ anti-PD-L1 + BCH vs anti-PD-L1, $^{**}P = 0.0038$ anti-PD-L1 + BCH vs BCH, $^{****}P < 0.0001$ anti-PD-L1 + BCH vs PBS + IgG by two-tailed unpaired two-way ANOVA. l, $n = 5$ in PBS + IgG, $n = 7$ in other groups biologically independent animals. The experiments were performed twice with similar results. TNFα: $^{*}P = 0.0478$ anti-PD-L1 + BCH vs PBS + IgG, $^{*}P = 0.0498$ anti-PD-L1 + BCH vs BCH, $^{*}P = 0.0425$ anti-PD-L1 + BCH vs anti-PD-L1 by two-tailed unpaired *t*-test. IFNγ: $^{*}P = 0.0269$ anti-PD-L1 + BCH vs PBS + IgG, $^{**}P = 0.0013$ anti-PD-L1 + BCH vs BCH, $^{**}P = 0.0331$ anti-PD-L1 + BCH vs anti-PD-L1 by two-tailed unpaired *t*-test. m, RNA-seq analysis showed expression of SLC43A2 transcripts in tumours and paired adjacent normal tissues in several types of tumour. CHOL, $n = 9$, $^{**}P = 0.0013$; ESCA, $n = 11$, $^{**}P = 0.0070$; HNSC, $n = 43$, $^{**}P = 0.0016$; KICH, $n = 24$, $^{****}P < 0.0001$; LIHC, $n = 50$, $^{*}P = 0.0273$ by two-tailed paired *t*-test.

Extended Data Figure 1a–c, $n = 4$ biologically independent samples. The experiments were performed three times with similar results. a, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. b, $^{***}P = 0.0002$ 100 μM vs FM, $^{***}P = 0.0005$ 50 μM vs FM, $^{***}P = 0.0019$ 30 μM vs FM by two-tailed unpaired *t*-test. c, $^{*}P = 0.0372$, $^{****}P < 0.0001$ by two-tailed unpaired *t*-test. d, e, $n = 4$ biologically independent samples. The experiments were performed twice with similar results. d, $^{*}P = 0.038$ sup vs FM, $^{*}P = 0.0139$ sup vs sup + met by two-tailed unpaired *t*-test. e, TNFα: $^{***}P = 0.0003$ sup vs FM, $^{****}P < 0.0001$ sup vs sup + met by two-tailed unpaired *t*-test. IFNγ: $^{****}P < 0.0001$ svs FM, $^{***}P = 0.0003$ sup vs sup + met by two-tailed unpaired *t*-test. f, g, $n = 11$ biologically independent donors in healthy control, and $n = 14$ biologically independent donors in patients with ovarian cancer. The experiments were performed once. g, $^{*}P = 0.0201$ by two-tailed unpaired *t*-test. h, $n = 3$ biologically

independent samples. The experiments were performed once. *P = 0.0531 by two-tailed unpaired t-test. i, n = 3 biologically independent samples. The experiments were performed twice with similar results. *P = 0.0306, **P = 0.0090 by two-tailed unpaired t-test. k–m, n = 5 biologically independent samples. The experiments were performed twice with similar results. ****P < 0.0001 by two-tailed unpaired t-test. o, p, n = 4 biologically independent animals. The experiments were performed twice with similar results. EC50 was determined by nonlinear regression (log (agonist) vs. response).

Extended Data Figure 2a–e, n = 4 biologically independent samples. RNA-seq in $CD8^+$ T cells. f, n = 4 biologically independent samples. The experiments were performed twice with similar results. h, i, n = 4 biologically independent samples. The experiments were performed twice with similar results. h, ****p sup vs FM, **P = 0.0097 sup vs sup + met by two-tailed unpaired t-test. i, **P = 0.0040 sup vs FM, **P = 0.0075 sup vs sup + met by two-tailed unpaired t-test. j, k, The experiments were performed three times with similar results.

Extended Data Figure 3a, b, The experiments were performed three times with similar results. c–e, n = 4 biologically independent samples. Gene signature comparison between *Dot1l*$^{-/-}$ and Dot1l$^{+/+}$ $CD8^+$ T cells through RNA array. f, n = 4 biologically independent samples. The experiments were performed twice with similar results. TNFα: *P = 0.0254 sup by two-tailed unpaired t-test. IFNγ: **P = 0.0032 by two-tailed unpaired t-test. Granzyme B: ***P = 0.0006 by two-tailed unpaired t-test. g, n = 5 biologically independent animals. The experiments were performed once. Day 16 *P = 0.0246 *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$, ****P < 0.0001 *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$ + anti-PD-L1, NS P = 0.9402 *Dot1l*$^{-/-}$ vs *Dot1l*$^{-/-}$ + anti-PD-L1 by two-tailed unpaired two-way ANOVA. h, n = 5 biologically independent samples. The experiments were performed twice with similar results. Day 14 **P = 0.003 by two-tailed unpaired two-way ANOVA. i, n = 4 biologically independent samples. The experiments were performed twice with similar results. *P = 0.0199 in dLN CD8, and *P = 0.0142 in tumour CD8 by two-tailed unpaired t-test. j, n = 4 biologically independent samples. The experiments were performed twice with similar results. Stat5a: *P = 0.0111 in fresh *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$ $CD8^+$ T cells, *P = 0.0144 in activated *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$ $CD8^+$ T cells. Stat5b: *P = 0.0116 in fresh *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$ $CD8^+$ T cells, ****P < 0.0001 in activated *Dot1l*$^{+/+}$ vs *Dot1l*$^{-/-}$ $CD8^+$ T cells. k, n = 3 biologically independent samples. The experiments were performed twice with similar results. Stat5b: *P = 0.0140 sup vs FM, **P = 0.0087 sup vs sup + met. l–n, n = 3 biologically independent samples. RNA-seq showed the effect of DOT1L inhibitor (SGC0946) on human $CD8^+$ T cells (PDB: GSE108694).

Extended Data Figure 4a, n = 6 biologically independent samples. **P = 0.0036 by two-tailed unpaired t-test. b, n = 6 biologically independent samples. **P = 0.0016 by two-tailed unpaired t-test. c, n = 5 biologically independent samples. ****P < 0.0001 by two-tailed unpaired t-test. d, n = 5 biologically independent samples. P = 0.0666 by two-tailed unpaired t-test. e, n = 6 biological independent clinical samples. The experiment was performed once. f, n = 6 biologically independent samples. *P = 0.0483 by two-tailed unpaired t-test. g, n = 6 biologically independent samples. **P = 0.0446 by two-tailed unpaired t-test. h, n = 5 biologically independent samples. **P = 0.0055 by two-tailed unpaired t-test. i, n = 5 biologically independent samples. P = 0.0596 by two-tailed unpaired t-test. j, n = 4 biologically independent samples. ****P < 0.0001 by two-tailed unpaired t-test. k, n = 4 biologically independent samples. ****P < 0.0001 by two-tailed unpaired t-test. l, n = 4 biologically independent samples. *P = 0.0155 by two-tailed unpaired t-test. m, n = 4 biologically independent samples. ****P < 0.0001 by two-tailed unpaired t-test. n, n = 5 biologically independent samples. NS P = 0.7891 in $CD45^-$ tumour, *P = 0.0217 in T cell by two-tailed unpaired t-test. o, n = 10 biologically independent samples. **P = 0.0045 by two-tailed unpaired t-test. p, n = 10 biologically independent animals. ****P < 0.0001 by two-tailed unpaired two-way ANOVA. q, n = 10 biologically independent samples. **P = 0.0011 met + anti-PD-L1 vs PBS + IgG, *P = 0.0219 met + anti-PD-L1 vs met + IgG, *P = 0.0492 met + anti-PD-L1 vs PBS + anti-PD-L1 by two-tailed unpaired t-test. r, n = 10 biologically independent samples. *P = 0.0220 met + anti-PD-L1 vs PBS + IgG, *P = 0.0444 met + anti-PD-L1 vs PBS + anti-PD-L1 by two-tailed unpaired t-test.

Extended Data Figure 5a, n = 3 biologically independent samples. The experiments were performed three times with similar results. **P = 0.0067, ****P < 0.0001 by two-tailed unpaired t-test. b, n = 3 biologically independent samples. The experiments were performed three times with similar results. ***P = 0.0002, ****P < 0.0001 by two-tailed unpaired t-test. c, n = 3 biologically independent samples. The experiments were performed twice with similar results. ****P < 0.0001 by two-tailed unpaired t-test. d, The experiments were performed twice with similar results. e, The experiments were performed three times with similar results. f, The experiments were performed twice with similar results. g, n = 3 biologically independent samples. The experiments were performed once. *P = 0.0277, **P = 0.0065 by two-tailed unpaired t-test. h, n = 5 biologically independent animals. The experiments were performed once. i, n = 7 biologically independent animals. The experiments were performed once. j, n = 5 biologically independent samples. The experiments were performed twice with similar results. *P = 0.0263 by two-tailed unpaired t-test. k, n = 8 biologically independent animals. The experiments were performed once. ****P < 0.0001 by two-tailed unpaired two-way ANOVA. l, The experiments were performed twice with similar results. m, n = 9 biologically independent animals. The experiments were performed once. n, n = 5 biologically independent animals. The experiments were performed twice with similar results. Day 35 *P = 0.0102 by two-tailed unpaired two-way ANOVA. o, n = 4 biologically independent samples. The experiments were performed twice with similar results. *P = 0.0415 in ID8 ascites, **P = 0.0389 in ID8 tumour by two-tailed unpaired t-test. p, n = 5 biologically independent samples. The experiments were performed twice with similar results. T cells: **P = 0.0029 BCH + anti-PD-L1 vs PBS + IgG, **P = 0.0094 BCH + anti-PD-L1 vs BCH, *P = 0.0461 BCH + anti-PD-L1 vs anti-PD-L1 by two-tailed unpaired t-test. CD8 cells: **P = 0.0049 BCH + anti-PD-L1 vs PBS + IgG, *P = 0.0136 BCH + anti-PD-L1 vs BCH, *P = 0.0480 BCH + anti-PD-L1 vs anti-PD-L1 by two-tailed unpaired t-test. q–s, Kaplan–Meier survival curves showed the relationship between levels of SLC43A2 and survival of patients with different types of tumour: cholangiocarcinoma (CHOL, p), low grade glioma (LGG, q), and lung squamous cell carcinoma (LUSC, r). The raw data were from TCGA. t–y, n = 12 independent patients. Single cell RNA-seq analyses were based on GSE72056. t, *P = 0.0485 by two-tailed paired t-test. v, Correlation was analysed using Pearson correlation analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA sequencing data that support the findings of this study have been deposited in NCBI Gene Expression Omnibus (GEO) under accession number GSE150887. All other data that supported the findings of this study are available from the corresponding author upon request. Source data are provided with this paper.

37. Roby, K. F. et al. Development of a syngeneic mouse model for events related to ovarian cancer. *Carcinogenesis* **21**, 585–591 (2000).
38. Wang, W. et al. Effector T cells abrogate stroma-mediated chemoresistance in ovarian cancer. *Cell* **165**, 1092–1105 (2016).
39. Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
40. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
41. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 63 (2016).

42. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theor. den Biowissenschaften* **131**, 281–285 (2012).
43. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med*. **50**, 96 (2018).
44. Lee, H.-J., Kremer, D. M., Sajjakulnukit, P., Zhang, L. & Lyssiotis, C. A. A large-scale analysis of targeted metabolomics data from heterogeneous biological samples provides insights into metabolite dynamics. *Metabolomics* **15**, 103 (2019).
45. Yuan, M. et al. Ex vivo and in vivo stable isotope labelling of central carbon metabolism and related pathways with analysis by LC-MS/MS. *Nat. Protocols* **14**, 313–330 (2019).

Acknowledgements We thank P. King for scientific input. We acknowledge support from the Advanced Genomics Core and Bioinformatics Core of the University of Michigan Medical School's Biomedical Research Core Facilities. This work was supported in part by research grants from the US NIH/NCI (W. Zou) (CA217648, CA123088, CA099985, CA193136 and CA152470) and the NCI Cooperative Human Tissue Network (CHTN). C.A.L. was supported by a 2017 AACR NextGen Grant for Transformative Cancer Research (17-20-01-LYSS) and an ACS Research Scholar Grant (RSG-18-186-01). Metabolomics studies performed at the University of Michigan were supported by NIH grant DK097153, the Charles Woodson Research Fund, and the University of Michigan Pediatric Brain Tumor Initiative. C.A.L. and W. Zou were supported by the NIH through the University of Michigan Rogel Cancer Center Grant (P30 CA046592).

Author contributions Y.B., W.L., and W. Zou proposed the research concept. Y.B. performed the majority of the experiments and explored the concept for the SLC transporter. Y.B., W.L., and W. Zou designed the experiments. W.L. and J.C. performed some in vivo experiments with *Dot1l*$^{-/-}$ mice. D.M.K., P.S., L.Z., Z.C.N. and C.A.L. designed, performed, and analysed the MS experiments for metabolite tests and analysis. S.L., J.L., M.C. and A.M.C. assisted with the RNA-seq and single cell RNA-seq data analysis. H.X., P.L., J.Y., L.V., W.S., and I.K. aided in mouse and human sample collection and FACS data analysis. S.W. and S.G performed mouse genotyping and breeding. J.R.L. and K.M. assisted in clinical study design and collection of specimens from patient with ovarian cancer. A.C., A.P., W. Zgodziński, G.W., I.W., and K.O. performed the clinical study on patients with colorectal cancer. Y.B., W.L., and W. Zou wrote the manuscript.

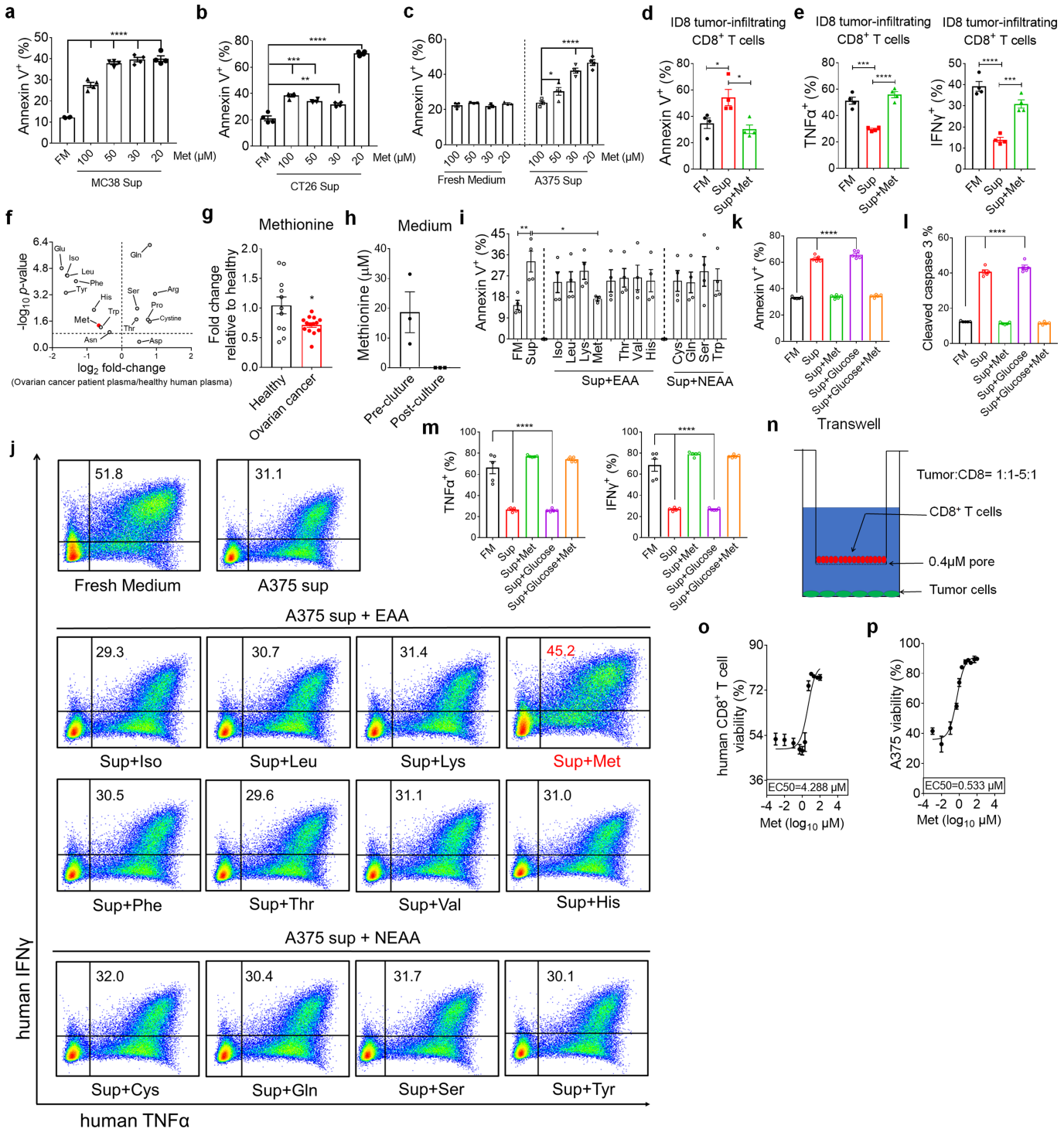
Competing interests The authors declare no competing interests.

Additional information

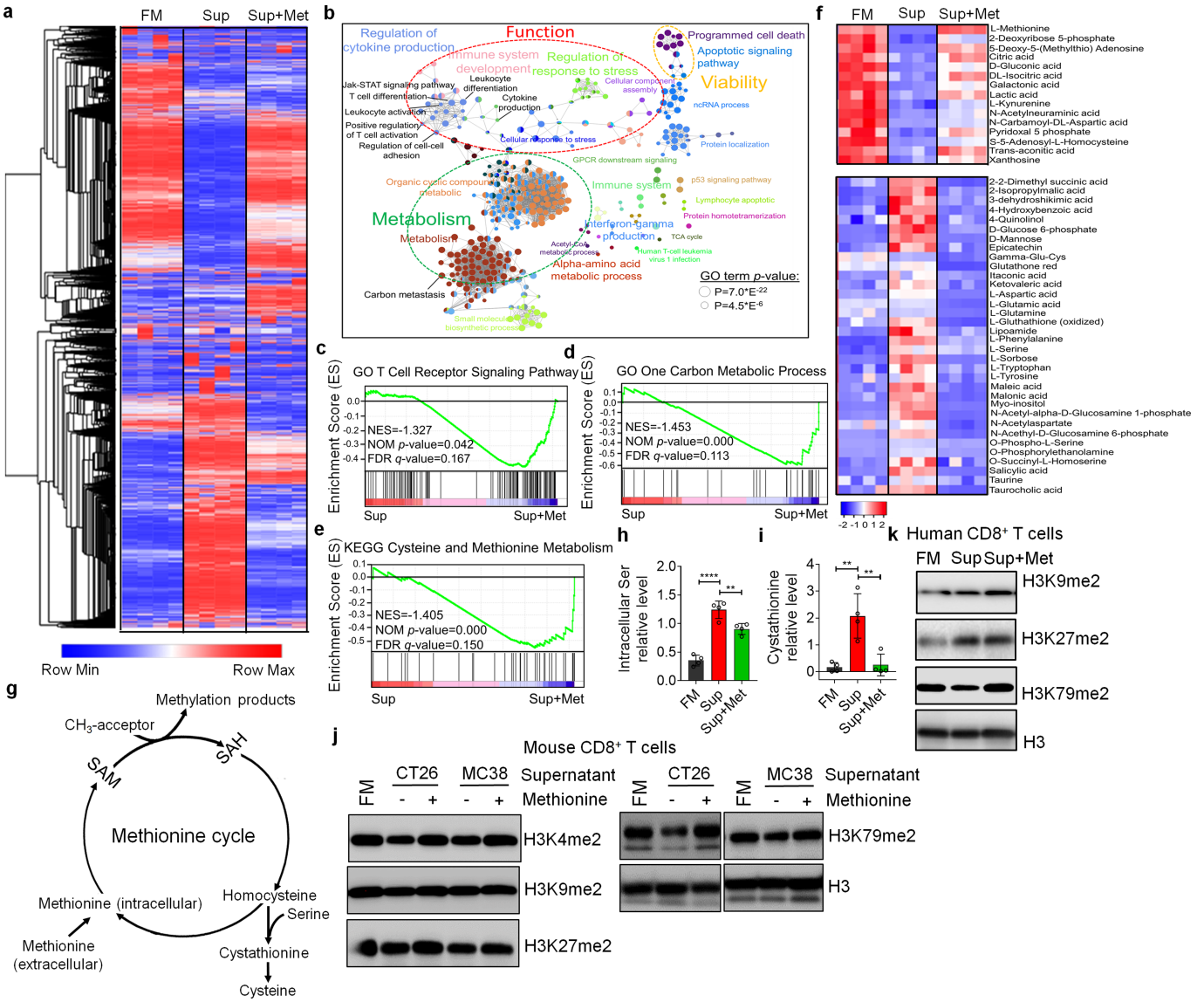
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2682-1.

Correspondence and requests for materials should be addressed to W.Z.

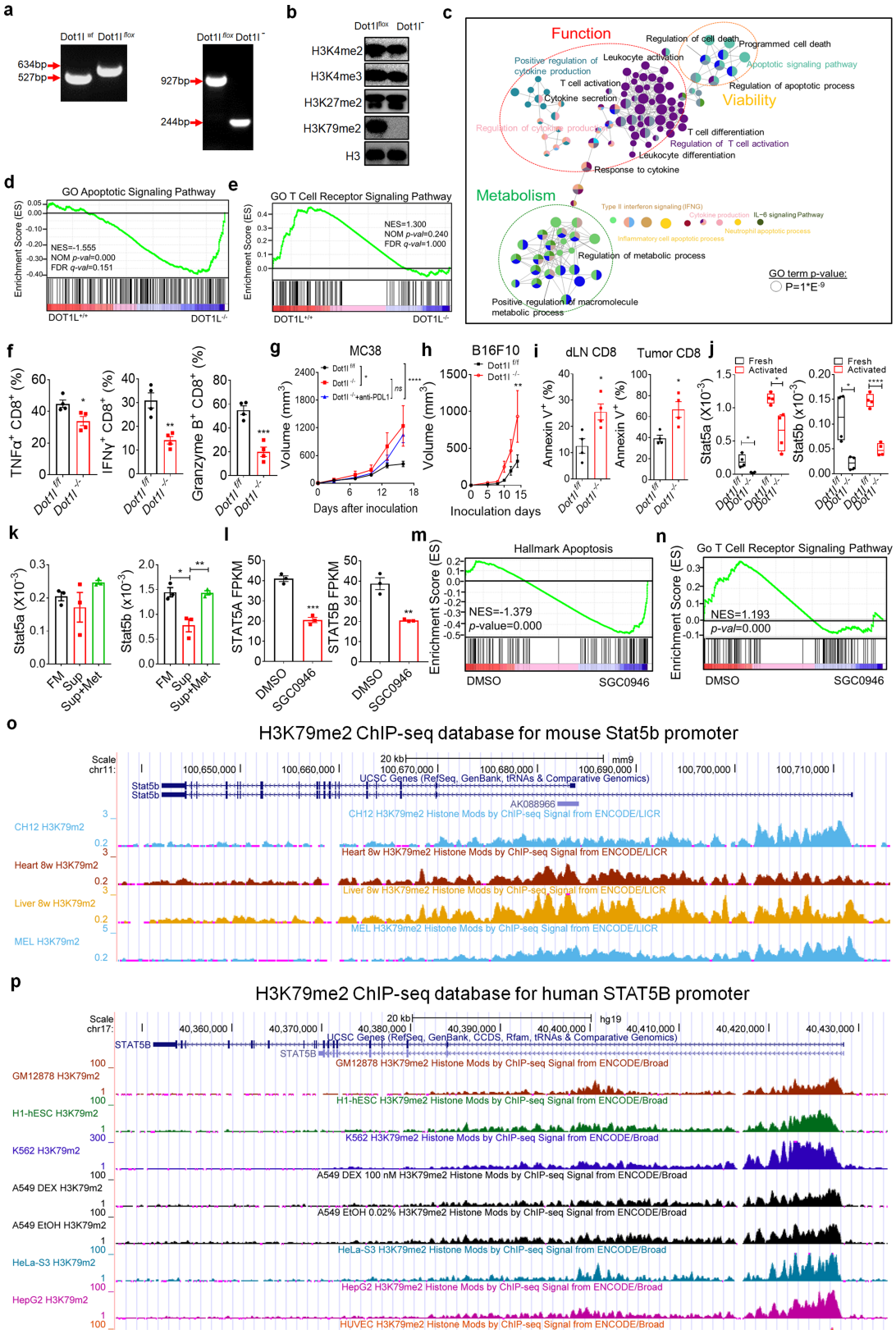
Peer review information *Nature* thanks Tak Mak, Stefani Spranger and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Tumour cells outcompete T cells for methionine to impair T cell function. a–**c**, Effect of tumour cells on T cell apoptosis. Tumour supernatants were collected from MC38 (**a**), CT26 (**b**), and human melanoma A375 (**c**) tumour cells cultured for 48 h with media containing different concentrations of methionine (Met). Then, $CD8^+$ T cells were cultured for 36 h with these tumour supernatants (Sup) or fresh medium. Apoptosis was determined by Annexin V staining. **d**, **e**, Effect of methionine on ID8 tumour infiltrating cells. T cells were cultured with fresh medium, ID8 supernatant (Sup), and supernatant plus methionine (Sup + Met). T cell apoptosis (**d**) and cytokine production (**e**) were determined by FACS. **f**, **g**, Amino acid levels in ovarian cancer patient plasma. Amino acids were detected in healthy donor and ovarian cancer patient plasma by liquid chromatography mass spectrometry (LC-MS). (**f**) Volcano showed plasma free amino acid changes. Red dot showed methionine (Met). (**g**) Plasma methionine in ovarian cancer patients vs healthy controls. **h**, Methionine concentration in pre- and post- tumour cultured medium. **i**, **j**, Effect of amino acid supplementation on human T cell function. $CD8^+$ T cells were cultured with A375 supernatants (Sup) supplemented with different amino acids for 36 h. FACS analysis showed T cell apoptosis (**i**) and effector cytokines (**j**). **k**–**m**, Effect of glucose supplementation on the role of methionine-affected T cell apoptosis and function. **n**, Schematic figure showing tumour and T cell co-culture in the Transwell system. **o**, **p**, Effect of methionine on human $CD8^+$ T cell (**o**) and tumour cell (**p**) viability, EC_{50} was determined by nonlinear regression (log(agonist) vs. response). Sup, tumour supernatant. Data are mean ± s.e.m. Information on sample sizes, experimental number, times, biological replicates, statistical tests, and *P* values is available in 'Statistics and reproducibility' (Methods).

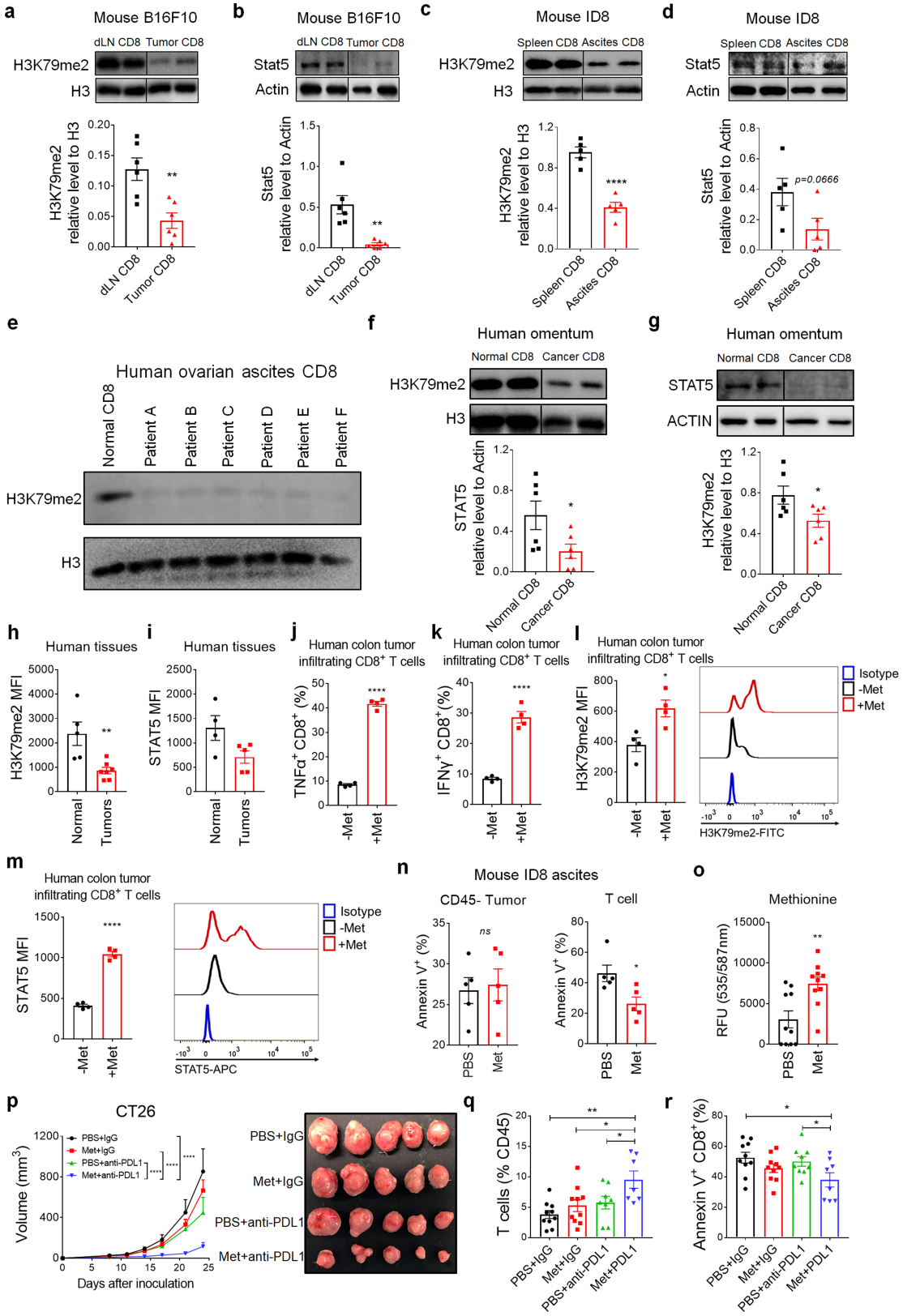


Extended Data Fig. 2 | Tumour alters CD8$^+$ T cell methionine metabolism to diminish H3K79me2. a, Gene profile changes in CD8$^+$ T cells. Mouse CD8$^+$ T cells were cultured with fresh medium, B16F10 tumour supernatant (Sup), or tumour supernatants plus methionine (Sup + Met) for 36 h. Gene profile changes were analysed by RNA-seq. **b**, Gene signatures were compared between groups from fresh medium and Sup. Functionally grouped network of enriched categories was generated for the hub genes and their regulators using ClueGO. Visualization has been carried out using Cytoscape 3.7.1. **c**–**e**, GSEA plot showed recovery of TCR signalling pathway (**c**) and methionine metabolism signalling (**d**, **e**) in CD8$^+$ T cells cultured with Sup + Met compared to Sup. **f**, Metabolites changes in CD8$^+$ T cells cultured with fresh medium, Sup and Sup + Met. Upper panel: Metabolites induced upon methionine supplementation. Lower panel: Metabolites suppressed upon methionine supplementation. **g**, The diagram of methionine cycle is shown. **h**, **i**, CD8$^+$ T cells were cultured with fresh medium, Sup, or Sup + Met for 36 h. Metabolites related to the methionine cycle, including intracellular serine (**h**) and L-cystathionine (**i**), were detected by MS. **j**, **k**, Effect of tumour supernatants on CD8$^+$ T cell histone methylation. Mouse (**j**) or human (**k**) CD8$^+$ T cells were cultured with or without methionine (Met) for 36 h with fresh medium, CT26 and MC38 tumour supernatants (**j**), or human A375 tumour supernatants (Sup) (**k**). T cell histone marks were determined by western blots. Data are mean ± s.e.m. Information on sample sizes, experimental number, times, biological replicates, statistical tests, and *P* values is available in 'Statistics and reproducibility' (Methods).



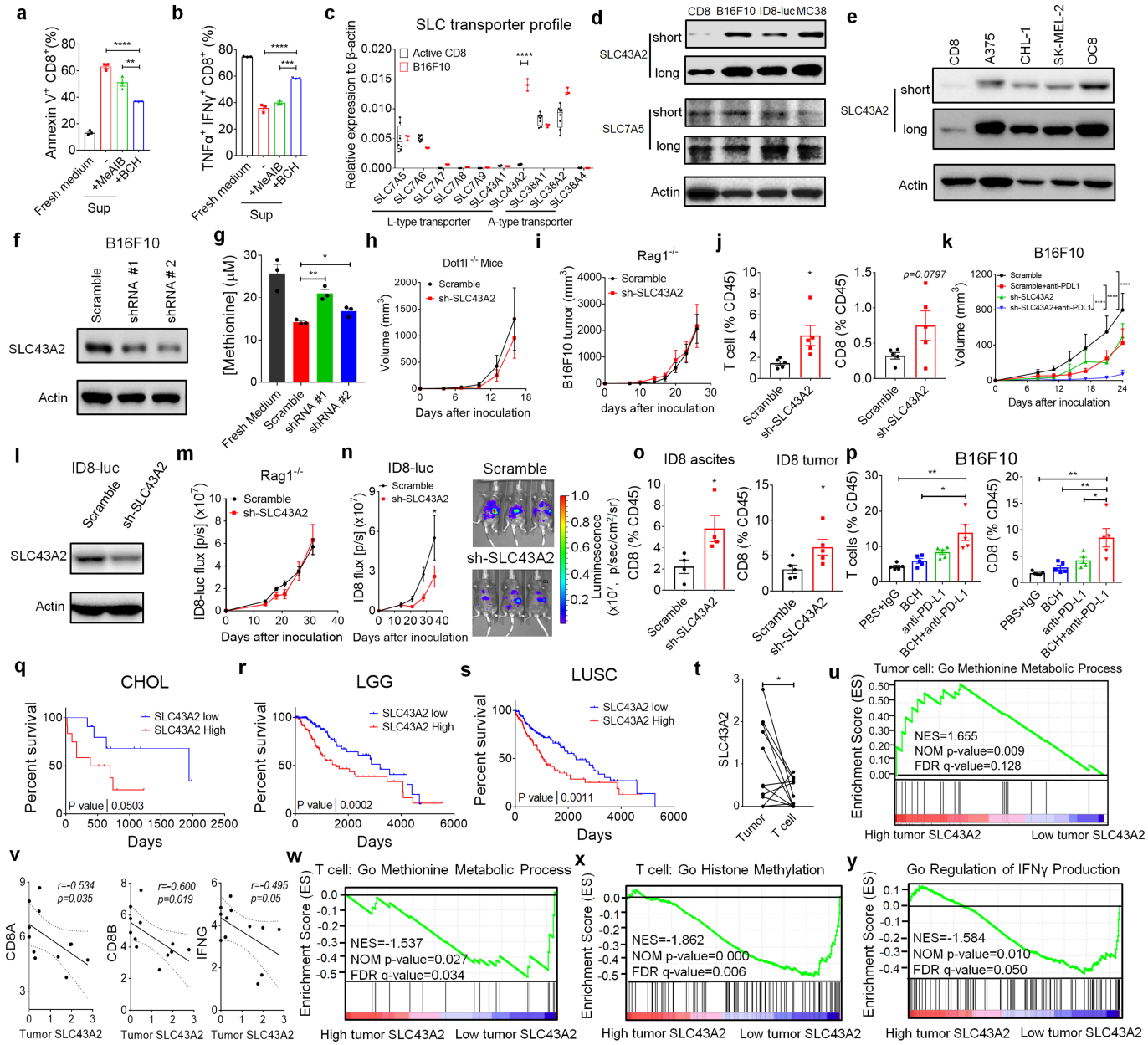
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Loss of H3K79me2 impairs T cell anti-tumour immunity through STAT5. a, Genotyping for *Dot1l*$^{f/f}$ and *Dot1l*$^{-/-}$ mice by PCR. **b**, Effect of Dot1l knockout on histone marks in T cells. **c**-**e**, Gene signature comparison between *Dot1l*$^{-/-}$ and *Dot1l*$^{f/f}$ CD8$^+$ T cells. Functionally grouped network of enriched categories was generated for the hub genes and their regulators using ClueGO. Visualization has been carried out using Cytoscape 3.7.1. (**c**). GSEA plot showed enriched apoptotic gene pathway (**d**) and impaired TCR signalling pathway (**e**) in *Dot1l*$^{-/-}$ CD8$^+$ T cells. **f**, Effect of DOT1L deficiency on T cell function in MC38 tumour. MC38 cells were inoculated into *Dot1l*$^{f/f}$ and *Dot1l*$^{-/-}$ mice. Expression of TNFα, IFNγ, and granzyme B in tumour infiltrating CD8$^+$ T cells was determined by FACS. **g**, Effect of anti-PD-L1 on tumour growth in *Dot1l*$^{+/+}$ and *Dot1l*$^{-/-}$ mice. **h**, **i**, B16F10 cells were inoculated into *Dot1l*$^{-/-}$ and *Dot1l*$^{f/f}$ mice. Effect of T cell DOT1L deficiency on tumour growth (**h**) and T cell viability (**i**) were monitored. **j**, Real-time PCR showed *Stat5a* and *Stat5b* transcripts in fresh or anti-CD3/CD28 activated *Dot1l*$^{f/f}$ and *Dot1l*$^{-/-}$ CD8$^+$ T cells. **k**, Real-time PCR showed *Stat5a* and *Stat5b* transcripts in activated CD8$^+$ T cells cultured with fresh medium, B16F10 tumour supernatants (Sup), or supernatants plus methionine (Sup + Met) for 24 h. **l**–**n**, RNA-seq showed the effect of DOT1L inhibitor (SGC0946) on human CD8$^+$ T cells (Database: GSE108694). STAT5A and STAT5B (**l**) transcripts were quantified in human CD8$^+$ T cells treated with DOT1L inhibitor SGC0946. GSEA enrichment plot showed enrichment of apoptotic gene pathway (**m**) and defects in T cell receptor related pathways (**n**) in human CD8$^+$ T cells treated with DOT1L inhibitor. **o**, **p**, H3K79me2 ChIP-seq in ENCODE database showing *Stat5b* promoter in mice (**o**) and humans (**p**). Data are mean ± s.e.m. Information on sample sizes, experimental number, times, biological replicates, statistical tests, and *P* values is available in 'Statistics and reproducibility' (Methods).

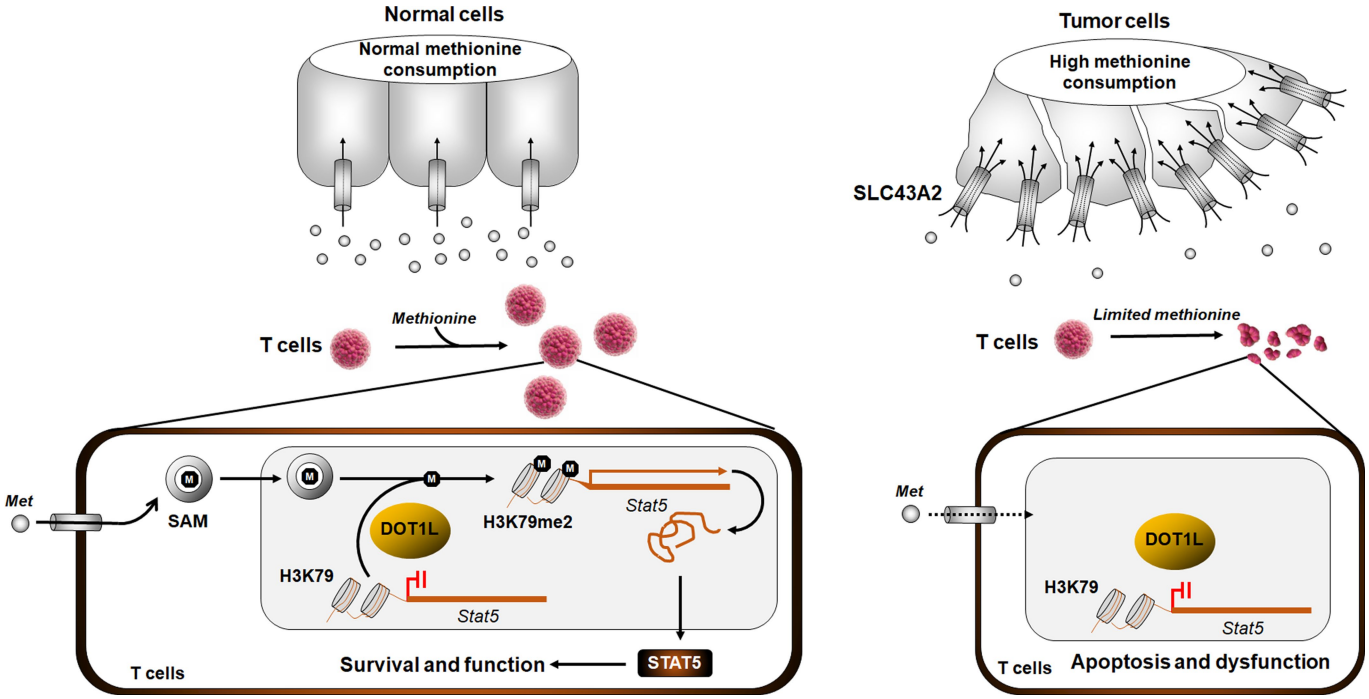


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Methionine supplementation promotes T cell anti-tumour immunity. a, b, H3K79me2 (**a**) and STAT5 (**b**) levels in $CD8^+$ T cells from tumour draining lymph node and tumour in B16F10 bearing mice. **c, d**, H3K79me2 (**c**) and STAT5 (**d**) levels in $CD8^+$ T cells from spleen and tumour ascites in ID8 bearing mice. **e**, H3K79me2 levels in $CD8^+$ T cells from healthy peripheral blood and human ovarian cancers ascites. **f**, **g**, H3K79me2 (**f**) and STAT5 (**g**) levels in $CD8^+$ T cells from healthy human blood and human ovarian cancer omentum tissues. **h, i**, FACS showed H3K79me2 and STAT5 levels in human tumour infiltrating $CD8^+$ T cells. **j**–**m**, Effect of methionine on human tumour infiltrating $CD8^+$ T cells. Human colorectal cancer infiltrating $CD8^+$ T cells were cultured with or without methionine. T cell cytokine production (**j**, **k**), H3K79me2 (**l**), and STAT5 (**m**) were analysed by FACS. One representative of four is shown. **n**, Effect of methionine supplementation on apoptosis of tumour infiltrating $CD8^+$ T cells and ID8 tumour cells in vivo. ID8 tumour bearing mice were treated with methionine or PBS. T cell and tumour cell apoptosis was determined by FACS. **o**, Methionine levels in ID8 tumour after methionine or PBS treatment. **p**–**r**, Effect of anti-PD-L1 on methionine-affected CT26 tumour progression. Mice bearing CT26 tumour were treated with anti-PD-L1, methionine, and their combination. Tumour volume (**p**), T cell tumour infiltration (**q**) and apoptosis (**r**) were assessed. Data are mean ± s.e.m. Information on sample sizes, experimental number, times, biological replicates, statistical tests, and *P* values is available in 'Statistics and reproducibility'.



Extended Data Fig. 5 | Tumour SLC43A2 correlates to poor T cell immunity. **a, b**, Effects of SLC inhibitors (BCH or MeAIB) on tumour cell affected CD8 T cell apoptosis (**a**) and cytokine production (**b**). **c**, Real-time PCR showed SLC transporter transcripts in activated $CD8^+$ T cells and B16F10 tumour cells. **d**, Western Blot showed SLC43A2 and SLC7A5 proteins in activated $CD8^+$ T cells and tumour cells. **e**, Western Blot showed SLC43A2 protein in human $CD8^+$ T cells and human tumour cells. **f**, Western Blot showed SLC43A2 knockdown efficiency in B16F10 cells. **g**, Effect of tumour cell SLC43A2 knockdown on methionine consumption. WT (scramble) and sh-SLC43A2 tumour cells were cultured with fresh medium containing 30 μM methionine for 24 h. Methionine concentration was measured by MS in fresh medium and supernatants. **h**, Wild-type and sh-SLC43A2 B16F10 tumour growth in *Dot1l*$^{-/-}$ mice. **i**, Wild-type and SLC43A2 knockdown B16F10 tumour growth in *Rag1*$^{-/-}$ mice. **j**, Effect of tumour SLC43A2 knockdown on T cell tumour infiltration in WT or sh-SLC43A2 B16F10 bearing mice. **k**, Effect of SLC43A2 knockdown and the combination of anti-PD-L1 on B16F10 bearing mice. **l**, Western Blot showed SLC43A2 knockdown efficiency in ID8-luc cells. **m**, Wild type and SLC43A2 knockdown ID8-luc tumour growth in *Rag1*$^{-/-}$ mice. **n**, **o**, Effect of tumour SLC43A2 knockdown on ID8 growth (**n**) and T cell tumour infiltration in WT or sh-SLC43A2 ID8 bearing mice. **p**, T cell tumour infiltration in B16F10 bearing mice treated with BCH, anti-PD-L1, or their combination. **q**–**s**, Kaplan–Meier survival curves showed the prognostic values of SLC43A2 expression in different types of tumour: Cholangiocarcinoma (CHOL, **q**), low grade glioma (LGG, **r**), and lung squamous cell carcinoma (LUSC, **s**). The raw data was from TCGA. **t**–**y**, The analysis was based on single cell RNA-seq data (GSE72056). **t**, SLC43A2 transcripts were compared in tumour cells versus tumour infiltrating T cells from the same human melanoma tissues. **u**, GSEA plots showed methionine metabolic process genes in tumour cells expressing high versus low SLC43A2. **v**, Correlation was analysed between CD8A, CD8B, IFNG transcripts in T cells and SLC43A2 transcripts in tumour cells in the same human melanoma tissues. **w**-**y**, GSEA enrichment plot analysis showed defective pathways in tumour infiltrating T cells in melanoma patients with high tumour SLC43A2 compared to low tumour SLC43A2. The pathways included T cell methionine metabolic process (**w**), histone methylation (**x**), and IFNγ production (**y**). Data are mean ± s.e.m. Information on sample sizes, experimental number, times, biological replicates, statistical tests, and *P* values is available in 'Statistics and reproducibility' (Methods).



Extended Data Fig. 6 | Graphical model. Model of how tumour cells outcompete T cells for methionine and disrupt T cell survival and function.

Extended Data Table 1 | ChIP primers for mouse *Stat5b*

Site	CHIP primer region	Primer Sequence
1	Stat5b promoter #1 (-186~-369)	F: 5'-TCATTCAGTCAGGATACGGGC-3'
		R: 5'-GAATTCCCCAGCTGAAAAGGC-3'
2	Stat5b promoter #2 (-790~-930)	F: 5'-AAAGGCGAAGAACAAACGGC-3'
		R: 5'-TACAAGTTCCGACCCACAGC-3'
3	Stat5b promoter #3 (-1072~-1240)	F: 5'-GCTTGAATGTGTGGTGGTGG-3'
		R: 5'-AGACAGCTCTCCTTCCGACT-3'
4	Stat5b promoter #4 (-1072~-1240)	F: 5'- CGTGCTCCTGCTGTCTAGAAGCTGGG-3'
		R: 5'- GGGATCGGCTCTGTCGGCGTC-3'
5	Stat5b promoter #5 (-3148~-3308)	F: 5'-AGGCCAGGAGTGTGTTTCTG-3'
		R: 5'-TGGAAATCAGCAGCTCTGGG-3'
6	Stat5b promoter #6 (-3886~-3940)	F: 5'-ATAGTGGGTGGCAGGGTTTG-3'
		R: 5'-CTGTCTACCTCATGGCGTCC-3'

Extended Data Table 2 | Characteristics of patients with colorectal cancer

Number	Gender	Age	Tumor histology	Grade	Primary tumor location	Stage	TNM
1	M	62	Adenocarcinoma tubulare	G2	Sigmoid colon	III	T4N1M0
2	M	84	Adenocarcinoma tubulare	G2	Transverse colon (Hepatic flexure)	II	T3N0M0
3	F	62	Adenocarcinoma tubulare	G2	Sigmo-rectal flexure	III	T3NxM0
4	M	66	Adenocarcinoma	G2	Ceacum	IV	T4N1M1
5	M	59	Adenocarcinoma tubulare	G3	Sigmoid colon	IV	T4N1M1
6	M	81	Adenocarcinoma tubulare	G2	Sigmoid colon	II	T3N0M0
7	F	65	Adenocarcinoma	G2	Sigmoid	III	T2N0M0

Extended Data Table 3 | Primers for RT–PCR and *Dot1l* mouse genotyping

Gene	Forward Sequence (5'-3')	Reverse Sequence (5'-3')
Slc3a2	GAGCGTACTGAATCCCTAGTCAC	GCTGGTAGAGTCGGAGAAGATG
Slc7a5	GGTCTCTGTTCACGTCCTCAAG	GAACACCAGTGATGGCACAGGT
Slc7a6	TCTACCTTCGCTGGAAAGAGCC	GCCACCAGAAACAAGGAGCAGA
Slc7a7	AAGGTGTTGGCGCTGATTGCAG	AGAGTGCCAGAGCAATGTCACC
Slc7a8	GCATACGTCACTGCAATGTCCC	GGAGCCATTGACTCCACCAAAC
Slc7a9	GGATTCCTCTGGTGACCGTATG	CAAGATGCTGGATAGAGAACGCG
Slc38a1	TACCAGAGCACAGGCGACATTC	ATGGCGGCACAGGTGGAACTTT
Slc38a2	GCGTTGGCATTCAATAGCACCG	TCGTAGATGGGAAGAACAGCGG
Slc38a4	CTCTTCACAGCAATGGCGTGGA	GACCTCAGGGTGGCAGACAAAA
Slc43a1	TTCCTGTGGAGCCTTGTCACCA	CTCCACCTTCTGTCTCTGCTCA
Slc43a2	CAGCATCCTTGAGTTCCTGGTC	TGATGTAGCCGATGACAGGAGC
Stat5a	CCTGTTTGAGTCTCAGTTCAGCG	TGGCAGTAGCATTGTGGTCCTG
Stat5b	CACAGTTCAGCGTCGGTGGAAA	CTGTGGCATTGTTGTCCTGGCT
Actb	CATTGCTGACAGGATGCAGAAGG	TGCTGGAAGGTGGACAGTGAGG
Dot1L-genotype-Dot1l alleles	GCCTACAGCCTTCATCATTC	GATAGTCTCAATAATCTCA
Dot1L-genotype-confirming excision	GAAGTTCCTATTCCGAAGTT	GAACCACAGGATGCTTCAG

nature research

Corresponding author(s): Weiping Zou

Last updated by author(s): May 8, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Flow cytometer BD LSRFortessa was used to run samples and data was acquired and analyzed by with FACSDiva v8.0.1 or FlowJo version 10. Epoch (BioTek) plate reader was used for requiring absorbance and data was analyzed by Gen5 software. The ID8-luc tumor growth was monitored by using the Xenogen IVIS Spectrum In Vivo Bioluminescence Imaging System (PerkinElmer). Real-time PCR was run on StepOnePlus system (Thermo fisher) and data was analyzed by StepOne Software v2.2.2. RNA-Seq data visualization has been carried out using Cytoscape v3.7.1. Data was also collected using standard software, such as Microsoft Excel 2016 and GraphPad Prism version 7.
Data analysis	GraphPad Prism version 7 was used for data analysis. FACS data were analyzed with FACSDiva v8.0.1 or FlowJo version 10. Functionally grouped network of enriched categories was generated for the hub genes and their regulators using ClueGO. RNA-Seq data visualization has been carried out using Cytoscape v3.7.1. Pathways gene enrichment were analyzed by using GSEA v4.0.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data supporting the results of this study are available from the corresponding author upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to calculate sample size. Sample size was determined to be adequate based on the magnitude and consistency of measurable differences between groups. The size of animal studies is between 5 to 20, which are commonly used in similar studies in literatures.
Data exclusions	No data was excluded for all in vitro experiments. No data was excluded for all in vitro experiments. We did not perform any pre-established exclusions for in vivo experiments. In in vivo survival studies we excluded small number of animals died due to other than tumor-related condition.
Replication	As reported in the figure legends, experiments were performed at least three times with similar results, the findings were reliably reproduced.
Randomization	For all in vivo experiments, animals were randomly assigned into a treatment group after tumor inoculation. The starting tumor burden in the treatment and control groups was similar before treatment.
Blinding	Investigators were not blinded to mouse genotypes during experiments. Tumor measurements were performed by person blinded to which animal was being measured.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*

Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies used for in vivo experiments, anti-mouse PD-L1 (Clone: 10F.9G2, Catalog # BE0101) and rat IgG2b isotype (Clone: LTF-2, Catalog # BE0090) were from Bioxcell validated in our previous works (https://www.jci.org/articles/view/96113 and https://www.nature.com/articles/s41586-019-1170-y#Sec2). Antibodies for functional studies: anti-human CD3 (Clone HIT3α, BD Biosciences, Catalog No. 555336, Working concentration: 5µg/L) and anti-human CD28 (Clone CD28.2, BD Biosciences, Catalog No. 555725, Working concentration: 2.5µg/L), anti-mouse CD3 (Clone 145-2C11, BD Biosciences, Catalog No. 553057, Working concentration: 5µg/L) and anti-mouse CD28 (Clone 37.51, BD Biosciences, Catalog No. 553294, Working concentration: 2.5µg/L).

Antibodies used for FACS: anti-mouse CD45 (30-F11,Thermo Fisher Scientific, Catalog # MCD4517), anti-mouse CD90 (53-2.1, Thermo Fisher Scientific, Catalog # 11-0902-82), anti-mouse CD4 (RM4-5, Thermo Fisher Scientific, Catalog # 47-0042-82), anti-mouse CD8 (53-6.7, Thermo Fisher Scientific, Catalog # 56-0081-82), anti-mouse IL-2 (JES6-5H4, Thermo Fisher Scientific, Catalog # 17-7021-82), anti-mouse TNF (MP6-XT22, Thermo Fisher Scientific, Catalog # 25-7321-82), anti-mouse IFNγ (XMG1.2, BD Biosciences, Catalog No. 563773), anti-mouse Granzyme B (NGZB, Thermo Fisher Scientific, Catalog # 12-8898-82), PE-Cy™7 Mouse Anti-Human CD3 (Clone UCHT1, BD Biosciences, Catalog No. 563423), APC-Cy™7 Mouse Anti-Human CD8 (Clone RPA-T8, BD Biosciences, Catalog No. 557760), Pacific Blue-anti-human IFNγ (4S.B3, BD Biosciences, Catalog No. 564791), PerCP-Cy™5.5 Mouse Anti-Human IFN-γ (4S.B3, BD Biosciences, Catalog No. 560742), APC-anti-human TNF (MAb11, BD Biosciences, Catalog No. 562084), FITC-Mouse Anti-Human TNF (MAb11, BD Biosciences, Catalog No. 552889), PE Rat anti-human IL-2 (MQ1-17H12, BD Biosciences, Catalog No. 560709), Alexa Fluor® 647 Mouse anti-Human Granzyme B (GB11, BD Biosciences, Catalog NO. 560212), APC-anti-STAT5 (REA549, Miltenyi Biotec Inc., Order no: 130-108-873), 7AAD (BD Biosciences, Catalog No. 559925), FITC-Annexin V (BD Biosciences, Catalog No. 556419).

Antibodies used for immunoblot and ChIP: anti-Histone H3 (di methyl K4) antibody (Abcam, Cat# ab194678, 1:1000), anti-Histone H3 (tri methyl K4) antibody (Abcam, Cat# ab8580, 1:1000), anti-Histone H3 (di methyl K9) antibody (Abcam, Cat# ab176882, 1:1000), anti-Histone H3 (di methyl K27) antibody (Abcam, Cat# ab24684, 1:1000), anti-Histone H3 (di methyl K79) antibody (Abcam, Cat# ab3594, 1:1000), anti-Histone H3 (tri methyl K79) antibody (Abcam, Cat# ab2621, 1:1000), anti-Histone H3 Antibody (Cell Signaling Technology, Cat# 9715, 1:1000), anti- STAT1 (Cell Signaling Technology, Cat# 14994, 1:1000), anti-STAT3 (Cell Signaling Technology, Cat# 12640, 1:1000), anti- STAT5 (Cell Signaling Technology, Cat# 94205, 1:1000), anti-Phospho-STAT5 (Tyr694) (Cell Signaling Technology, Cat# 4322, 1:1000), Normal Rabbit IgG (Cell Signaling Technology, Cat# 2729, 1:1000), anti-β-Actin (Antibody Cell Signaling Technology, Cat# 4967, 1:1000), anti-SLC43A2 antibody (Abcam, Cat# ab107426, 1:1000), anti-SLC7A5 Polyclonal Antibody (Invitrogen, Cat# PA5-50485, 1:1000).

Validation

All antibodies for FACS and western blot were well-recognized clones in the field and validated by the manufacturers. These antibodies are further validated and routinely used in our lab.
Antibodies targeting SLC43A2 were validated by knockdown through MISSION shRNA (Sigma) and the GIPZ Lentiviral shRNA systems, and verification of the decrease of a band of the predicted molecular weight by immunoblotting.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Human cells (including A375, CHL-1, SK-MEL-2, 293T cells) and mouse tumor cells (including B16F10 and CT26 cells) were obtained from ATCC. Mouse ID8-luc and MC38 cells and human primary high grade serous ovarian carcinoma cells (OC8) are cited.

Authentication

STR fingerprint analysis

Mycoplasma contamination

All cell lines in our laboratory are routinely tested for mycoplasma contamination and cells used in this study are negative for mycoplasma.

Commonly misidentified lines
(See ICLAC register)

No cell line used in the paper is listed in ICLAC database.

Palaeontology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Six- to eight-week-old female wild-type C57BL/6, BALB/c, and Rag1 knock out (KO) mice were from the Jackson Laboratory. Dot1l flox/flox mice were bred to CD4-Cre mice to generate mice with specific DOT1L deletion in T cells. All mice were maintained under pathogen-free conditions.

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected from field.

Ethics oversight

University Committee on the Use and Care of Animals at the University of Michigan

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Colorectal cancer (stage from II to IV) patients were recruited for the methionine supplementation study. There are 5 males and 2 females, with the age range from 59-84 (average 68.4).
Recruitment	None selection had been made, every patient who meet the eligibility criteria were selected for study.
Ethics oversight	This study was conducted according to the Declaration of Helsinki and approved by the institutional review board, with written informed consent obtained from all patients.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For cell apoptosis analysis, cells were treated, collected and rinsed with 1x binding buffer, and then stained with Annexin V and 7AAD in binding buffer at 4°C for 10 mins and directly run on a flow cytometer. To quantify T cell and cytokine production, single-cell suspensions were prepared from fresh tumor tissues and lymphocytes were enriched by density gradient centrifugation. For cytokine staining, lymphocytes were incubated in culture medium containing PMA (5 ng/ml), Ionomycin (500 ng/ml), Brefeldin A (1: 1000) and Monensin (1: 1000) at 37°C for 4 hours. Anti-CD45 (30-F11), anti-CD90 (53-2.1), anti-CD4 (RM4-5) and anti-CD8 (53-6.7) were added for 20 minute for surface staining. The cells were then washed and resuspended in 1 ml of freshly prepared Fix/Perm solution (BD Biosciences) at 4°C for overnight. After being washed with Perm/Wash buffer (BD Biosciences), the cells were stained with anti-IL-2 (JES6-5H4), anti-TNFa (MP6-XT22), anti-IFNg (XMG1.2) and anti-granzyme B (NGZB) for 30 min. For STAT5, cells were stained with APC-anti-STAT5 (REA549, Miltenyi Biotec Inc., Bergisch Gladbach, Germany). For DOT1L and H3K79me2 intracellular staining, the cells were first stained with DOT1L or H3K79me2 antibodies (Abcam), and then stained using a FITC-conjugated goat anti-rabbit IgG (H+L) secondary antibody (Invitrogen).
Instrument	All samples were acquired on BD LSRFortessa.
Software	All data were analyzed with FACS DIVA software v. 8.0 (BD Biosciences), or FlowJo V10 (LLC).
Cell population abundance	When cells were sorted or enriched, the purity was confirmed by flow cytometry and in each case was above 90% purity.
Gating strategy	The cells were gated on FSC-A/SSC-A basis on the location known to contain lymphoid cells and tumor cells. To analyze cytokine production by mouse T cells, CD45+ CD90+ population was first gated, and CD8+ and CD4+ populations were then gated. In CD8 gate, the percentage of IL-2, TNFa+, IFNg+ or Granzyme B cells were analyzed. To analyze cytokine production by human T cells, CD3+ population was first gated, and CD8+ population were then gated. The percentage of IL-2, TNFa+, IFNg+ cells were then analyzed.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.

Article

Age-induced accumulation of methylmalonic acid promotes tumour progression

https://doi.org/10.1038/s41586-020-2630-0

Received: 7 May 2019

Accepted: 10 July 2020

Published online: 19 August 2020

Check for updates

Ana P. Gomes[1,2,11]✉, Didem Ilter[1,2,11], Vivien Low[1,3], Jennifer E. Endress[1,3], Juan Fernández-García[4,5], Adam Rosenzweig[1,2], Tanya Schild[1,3], Dorien Broekaert[4,5], Adnan Ahmed[1,6], Melanie Planque[4,5], Ilaria Elia[4,5], Julie Han[1,2], Charles Kinzig[1,7], Edouard Mullarky[1,8], Anders P. Mutvei[1,2], John Asara[9], Rafael de Cabo[10], Lewis C. Cantley[1,8], Noah Dephoure[1,6], Sarah-Maria Fendt[4,5] & John Blenis[1,2,6]✉

The risk of cancer and associated mortality increases substantially in humans from the age of 65 years onwards[1–6]. Nonetheless, our understanding of the complex relationship between age and cancer is still in its infancy[2,3,7,8]. For decades, this link has largely been attributed to increased exposure time to mutagens in older individuals. However, this view does not account for the established role of diet, exercise and small molecules that target the pace of metabolic ageing[9–12]. Here we show that metabolic alterations that occur with age can produce a systemic environment that favours the progression and aggressiveness of tumours. Specifically, we show that methylmalonic acid (MMA), a by-product of propionate metabolism, is upregulated in the serum of older people and functions as a mediator of tumour progression. We traced this to the ability of MMA to induce SOX4 expression and consequently to elicit transcriptional reprogramming that can endow cancer cells with aggressive properties. Thus, the accumulation of MMA represents a link between ageing and cancer progression, suggesting that MMA is a promising therapeutic target for advanced carcinomas.

Considering the growing body of evidence that cancer cell-extrinsic factors are key in modulating tumour progression, we hypothesized that ageing might produce a systemic environment that supports tumour progression and aggressiveness. To test this hypothesis, we cultured human cancer A549 and HCC1806 cells in 10% human serum (HS) from 30 young (age ≤30 years) and 30 old (age ≥60 years) healthy donors (Fig. 1a, Supplementary Table 1). Whereas the majority (25 out of 30) of cells treated with young donor serum maintained their epithelial morphology, cells treated with 25 out of the 30 old donor sera became mesenchymal, losing their polarity and displaying a spindle-shaped morphology (Extended Data Figs. 1–3). These phenotypes were independent of donor ethnicity, and resembled the epithelial-to-mesenchymal transition (EMT), a developmental process that is hijacked by cancer cells to acquire pro-metastatic properties[13]. Cells cultured with aged-donor serum displayed a pronounced loss of the epithelial marker E-cadherin and gain of the mesenchymal markers fibronectin and vimentin, in addition to increased expression of serpine1 and MMP2—proteins associated with aggressive phenotypes[13] (Fig. 1b, Extended Data Fig. 4a, b). Moreover, the aged sera promoted resistance to two distinct and widely used chemotherapeutic drugs, carboplatin and paclitaxel (Fig. 1c, Extended Data Fig. 4c). To determine whether the cells treated with the old donor sera would also show heightened metastatic potential, we treated MDA-MB-231 breast cancer cells with HS before injecting them into the tail veins of athymic mice. In contrast to the young sera, the aged sera robustly potentiated the ability of the cells to colonize the lungs and form metastatic lesions (Fig. 1d, e). These data show that systemic ageing and age-induced circulatory factors help to promote the acquisition of aggressive properties of cancers.

Pro-inflammatory factors have a key role in tumour progression[14], and also contribute to age-related diseases[15]. However, proteomic analysis of the old sera did not show a general pro-inflammatory signature that could explain the aggressive properties we observed in the cancer cells (Extended Data Fig. 4d). Considering the effectiveness of metabolic interventions such as diet, exercise and caloric restriction in mitigating susceptibility to and outcomes of cancer[1,9–12], we examined the metabolic compositions of the donor sera. Out of the 179 circulatory metabolites detected by targeted metabolomics, only 10 were altered at a statistically significant level (two-sided *t*-test, $P < 0.5$) (Supplementary Table 1). A pronounced decline in levels of glutathione, spermidine, glutamine and α-ketoglutarate was expected, considering their known or suggested roles in the ageing process[16–19] (Supplementary Table 1). Notably, only three metabolites were consistently increased in the sera of aged donors: phosphoenolpyruvate, quinolinate and methylmalonic acid (MMA) (Extended Data Fig. 4e). To test whether any of these three metabolites was responsible for inducing the pro-aggressive effects, we treated A549 cells with each metabolite. Only MMA induced a complete pro-aggressive EMT-like phenotype with a decline in E-cadherin

[1]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [2]Department of Pharmacology, Weill Cornell Medicine, New York, NY, USA. [3]The Biochemistry, Structural, Developmental, Cell and Molecular Biology Allied PhD Program, Weill Cornell Medicine, New York, NY, USA. [4]Laboratory of Cellular Metabolism and Metabolic Regulation, VIB-KU Leuven Center for Cancer Biology, VIB, Leuven, Belgium. [5]Laboratory of Cellular Metabolism and Metabolic Regulation, Department of Oncology, KU Leuven and Leuven Cancer Institute (LKI), Leuven, Belgium. [6]Department of Biochemistry, Weill Cornell Medicine, New York, NY, USA. [7]Weill Cornell Medicine/Rockefeller University/Sloan Kettering Tri-Institutional MD-PhD Program, New York, NY, USA. [8]Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [9]Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA. [10]Laboratory of Experimental Gerontology, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA. [11]Present address: Department of Molecular Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, USA. ✉e-mail: ana.gomes@moffitt.org; job2064@med.cornell.edu

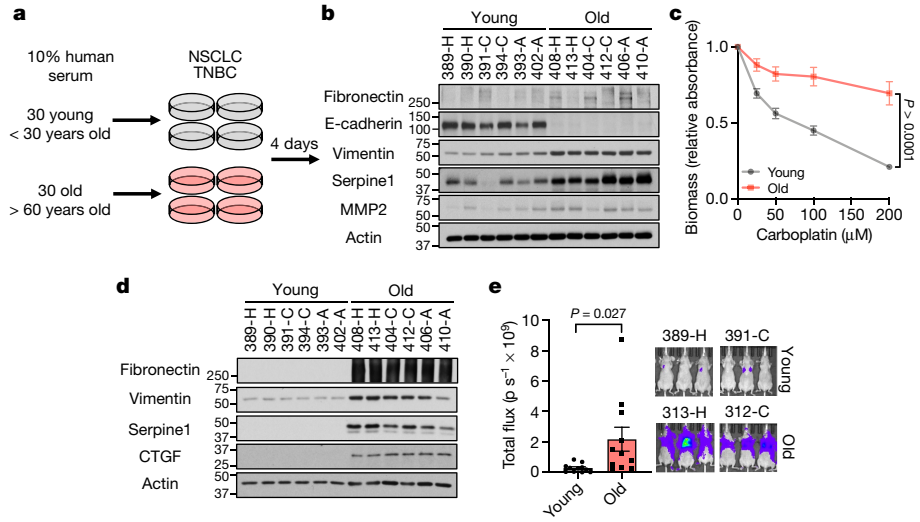


Fig. 1 | An age-induced circulatory factor promotes cancer aggression. **a**, Diagram showing experimental design (see Methods). **b**, Immunoblots of A549 cells cultured for 4 days in HS from young or old donors; see Extended Data Fig. 4a (total of $n = 30$ biologically independent samples per HS donor group). **c**, Resistance to carboplatin in A549 cells cultured for 4 days in HS ($n = 15$ biologically independent samples per HS donor group, two-sided ANOVA). **d**, **e**, Metastatic properties of MDA-MB-231-luciferase cells cultured for 5 days in HS evaluated by immunoblots (**d**; $n = 6$ biologically independent samples per HS donor group) and lung colonization assay (**e**; $n = 11$ biologically independent samples, each the average of three mice used as technical replicates, per HS donor group, two-sided t-test); examples shown to the right. **c**, **e**, Data presented as mean ± s.e.m. For gel source data, see Supplementary Fig. 2.

and a concurrent increase in fibronectin and vimentin (Extended Data Fig. 4f).

MMA is a dicarboxylic acid that is primarily a by-product of propionate metabolism. Propionyl-CoA, produced from catabolism of branched chain amino acids and odd chain fatty acids, yields succinyl-CoA in a manner that depends on vitamin B12 to fuel the tricarboxylic acid (TCA) cycle. The accumulation of MMA results from increased flux through and/or deregulation of the enzymes in this pathway and is a marker for a group of inborn metabolic diseases called methylmalonic acidaemias, as well as of vitamin B12 deficiency[20]. Large-scale exploratory metabolomic experiments are notorious for their lack of sensitivity and quantification[21]; therefore, to gain a deeper insight, we measured the absolute concentration of MMA in the sera from all 60 donors. This analysis revealed that MMA levels were higher in the sera of the old donors (15–80 μM) than in that of the young donors (0.1–1.5 μM) (Fig. 2a). Moreover, in the case of the ten outlier samples (five samples from old donors that did not induce EMT and five samples from young donors that did induce EMT), MMA levels consistently correlated with the phenotypes observed in cancer cells, supporting the idea that MMA is, at least in part, responsible for the observed age-related aggressive phenotypes (Extended Data Fig. 4g). Vitamin B12 levels are known to decline with age[22], and our measurements of vitamin B12 in the sera revealed a modest decline in old donors (Extended Data Fig. 4h); however, this decline did not correlate with MMA levels in the outlier samples (Extended Data Fig. 4i). Although we cannot exclude the possibility that vitamin B12 deficiency contributes to the accumulation of MMA with age, other factors such as deregulation of propionate metabolism in a major organ are also likely to be involved.

To better understand the pro-aggressive properties of MMA, we treated HCC1806, A549 and MCF-10A cells (a common model for EMT studies) with MMA. Concentrations of 1 mM and above were sufficient to induce an EMT-like phenotype and the expression of pro-aggressive proteins (Fig. 2b, Extended Data Fig. 5a, b). Notably, the pro-aggressive effects of MMA were specific, as different acids with similar structures and pK_a values did not induce the same complete phenotype under the specific conditions used (Extended Data Fig. 5c, d). MMA also induced resistance to carboplatin and paclitaxel (Extended Data Fig. 5e–h), increased the migratory and invasive capacity of the cells (Fig. 2c, Extended Data Fig. 5i), and promoted stem-like properties, as shown by an upregulation of CD44 and a decline in CD24 (Extended Data Fig. 5j, k). Treatment of MDA-MB-231 cells in vitro with MMA increased markers of aggressiveness (Extended Data Fig. 5l) and was sufficient to robustly increase the ability of the cells to colonize the lungs of athymic mice in a concentration-dependent manner (Fig. 2d). Together, these data support the idea that MMA promotes pro-aggressive traits and contributes to the cellular plasticity required for tumour progression.

Although MMA concentrations above 1 mM were required to potently induce aggressive traits in cancer cells in vitro, MMA concentrations measured in serum from old donors were much lower than this (Fig. 2a). Further analysis demonstrated that the intracellular concentrations of MMA achieved within 4 h of treatment with old serum or 5 mM MMA were substantially different (Extended Data Fig. 6a, b). In fact, 5 mM MMA took 48 h to produce intracellular concentrations similar to the ones observed after 4 h treatment with old serum. By contrast, a more cell-permeable version of MMA (dimethyl MMA) could induce pro-metastatic effects at concentrations as low as 10–50 μM (Extended Data Fig. 6c, d), suggesting that the discrepancy in concentrations is because added MMA has a lower cell permeability than endogenous MMA in donor serum. To assess whether another component of the serum could facilitate the entrance of MMA into cancer cells, we depleted the old serum of lipids or of molecules larger than 3 kDa—two manipulations that should not affect the levels of polar metabolites such as MMA. In both cases, the ability of the depleted old serum to induce pro-aggressive properties was abolished (Extended Data Fig. 6e). Strikingly, both manipulations also caused a pronounced decrease in serum MMA levels (Extended Data Fig. 6f), indicating that the MMA responsible for this phenotype is complexed with lipidic structures larger than 3 kDa in the serum that facilitate its entry into cancer cells. To test this hypothesis, we first complexed MMA with synthetic lipidic structures (lipofectamine) or with lipidic structures purified from fetal bovine serum (FBS). With both approaches, the concentration of MMA necessary to induce pro-aggressive properties was reduced to the levels similar to that of the old donor serum (Extended Data Fig. 6g–i). Moreover, MMA complexed with lipidic structures from FBS produced a similar intracellular concentration

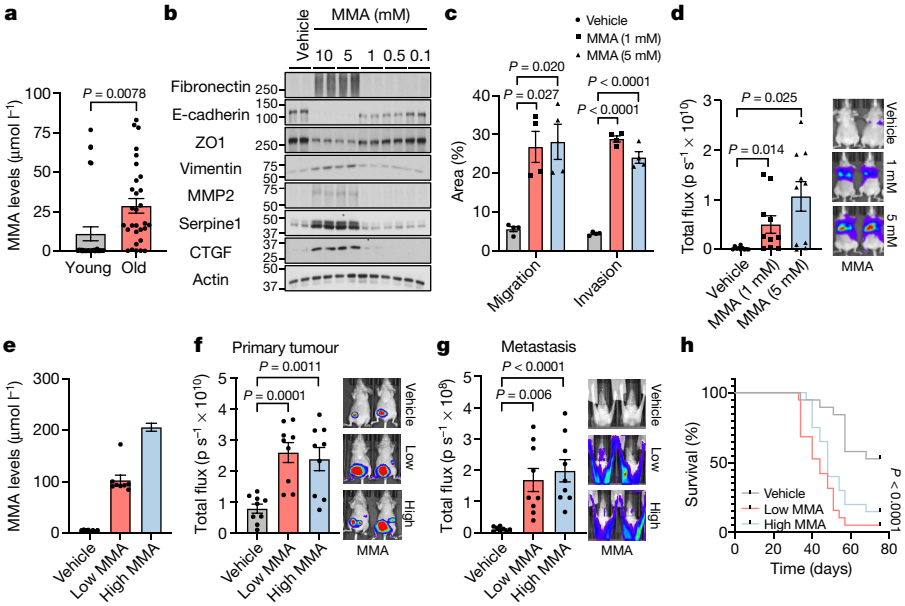


Fig. 2 | MMA induces aggressive traits of cancer cells. a, Concentrations of MMA in all HS samples ($n = 30$ biologically independent samples per HS donor group). **b**, Immunoblots of A549 cells treated with MMA for 10 days; representative images ($n = 4$ independent experiments). **c**, Transwell migration and invasion assays of MCF-10A cells treated with MMA for 10 days ($n = 4$ independent experiments). **d**, Lung colonization assay of MDA-MB-231-luciferase cells treated with MMA for 5 days ($n = 10$ mice per group; example mice shown to right). **e**–**g**, End-point serum MMA concentrations (**e**; $n = 8$ mice per group), bioluminescence intensity of the primary tumours (**f**, $n = 9$ mice per group), and metastases (**g**; $n = 9$ mice per group) in mice that were xenografted with MDA-MB-231-luciferase cells and subcutaneously injected with MMA daily. **h**, Kaplan–Meier curve of mice xenografted with MDA-MB-231-luciferase cells and treated with MMA either subcutaneously or through drinking water ($n = 19$ mice per experimental group). **a**, **c**–**g**, Mean ± s.e.m., two-sided *t*-test; **h**, Mantel–Cox test. For gel source data, see Supplementary Fig. 2.

of MMA within the same time frame as treatment with old donor serum (Extended Data Fig. 6j). In support of this idea, treatment of cancer cells with lipidic structures isolated from old serum, but not from young serum, or isolated from young serum and loaded with MMA at concentrations similar to the ones found in the old serum, was sufficient to drive pro-aggressive properties (Extended Data Fig. 6k). Conversely, depletion of lipidic structures from old serum resulted in a reduction in total serum MMA levels and was sufficient to abrogate the pro-aggressive phenotype (Extended Data Fig. 6l, m). Orthotopic injections of MDA-MB-231 cells into the mammary fat pads of athymic mice with elevated circulatory MMA levels (Fig. 2e, Extended Data Fig. 7a) further demonstrated that circulatory MMA has a substantial role in tumour progression by promoting tumour growth and metastatic spread; there was concomitant significant decrease in survival of this cancer model (Fig. 2f–h, Extended Data Fig. 7b, c). Our data show that MMA, complexed with lipidic structures, is a circulatory factor that contributes to the pro-aggressive effects of ageing in cancer cells and is sufficient to drive tumour progression and aggressiveness.

To investigate how MMA promotes the observed cellular plasticity, we performed a global transcriptomic analysis in A549 cells treated with MMA for 10 days. We found that MMA induced marked transcriptional reprogramming (Extended Data Fig. 7d, Supplementary Table 2). Gene set enrichment analysis (GSEA) showed that MMA regulates genetic programs associated with cell fate decisions, such as wound healing and pattern specification (Extended Data Fig. 7e), as well as genes involved in resistance to chemotherapeutic drugs, including several members of the ABC transporter family (Extended Data Fig. 7f). Many of the upregulated genes encode secreted proteins known to remodel the tumour microenvironment, including factors that promote reorganization of the extracellular matrix, immunosuppressive cytokines, and ligands that promote cell-to-cell communication (Extended Data Fig. 7g, h). Thus, MMA controls a panoply of genetic programs, remodelling both the tumour cells and the microenvironment to promote aggressiveness and cancer progression.

GSEA also showed that MMA positively regulates transcription (Extended Data Fig. 7e), suggesting that the observed pro-aggressive transcriptional reprogramming may be mediated through this role. To find the transcriptional regulators involved, we again performed global transcriptomic analysis at an earlier time point during MMA treatment (day 3). Unexpectedly, out of 439 induced genes (upregulated at least 1.5-fold), only 11 encoded transcription factors; 9 of these were significantly changed when validated with quantitative PCR (qPCR) (Extended Data Fig. 7i, Supplementary Table 2). One of the most upregulated of these transcription factor genes was *SOX4*. Notably, SOX4 is a marker of poor prognosis that contributes to tumour progression and metastasis formation, with aberrantly high expression in a wide variety of aggressive cancers[23–26], and is known to be a master regulator of EMT[27]. The levels of SOX4 were considerably increased in a variety of cell models treated with different concentrations of MMA, as well as in cells cultured with aged serum (Fig. 3a, b, Extended Data Fig. 7j–l), supporting the idea that SOX4 mediates the pro-aggressive phenotype observed. A comparison between genes induced by SOX4[28] and those induced by MMA treatment revealed a statistically significant overlap of 199 genes (Extended Data Fig. 7m, Supplementary Table 2). Functional annotation clustering analysis revealed that the overlapping genes—including *FN1*, *CDH2*, *MMP2*, *IL32* and *TGFB1I1*—are associated with pro-aggressive genetic programs. To better understand the relationship between MMA and SOX4, we used short hairpin RNA (shRNA) to suppress *SOX4* expression. When *SOX4* expression was suppressed, treatment with MMA failed to upregulate the mRNA levels of several of these genes (Extended Data Fig. 8a–d). Moreover, suppression of *SOX4* blocked the ability of MMA or old serum to induce EMT and aggressive markers (Extended Data Fig. 8e, f). Finally, SOX4 depletion fully abrogated the ability of MMA to promote migratory and invasive properties (Extended Data Fig. 8g) or resistance to chemotherapeutic drugs (Extended Data Fig. 8h, i), and the ability of MDA-MB-231-luciferase cells to form colonies in the lungs of athymic mice upon treatment with MMA or with old serum (Fig. 3c, d).

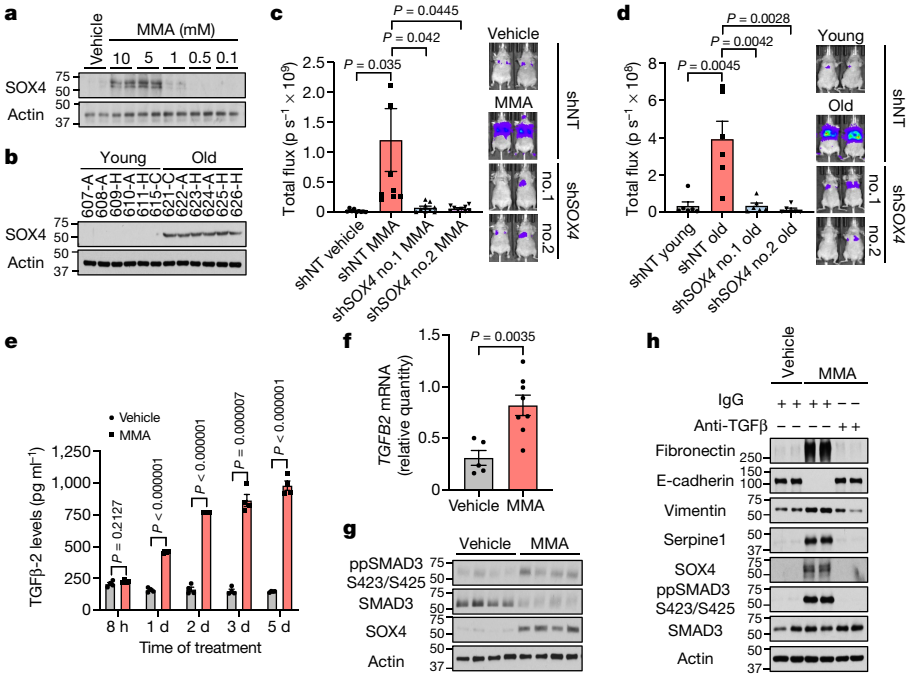


Fig. 3 | MMA triggers pro-aggressive transcriptional reprogramming by activation of TGFβ signalling and consequent induction of SOX4. **a**, **b**, Immunoblots of A549 cells treated with MMA for 10 days (**a**; $n = 4$ independent experiments) or HS for 4 days (**b**; $n = 6$ biologically independent samples, each the average of three mice used as technical replicates, per HS donor group). **c**, **d**, Lung colonization assay of MDA-MB-231-luciferase cells with SOX4 knockdown (sh*SOX4* no. 1 or no. 2) and treated with 5 mM MMA (**c**; $n = 8$ mice per group) or HS from old donors (**d**; $n = 6$ mice per group) for 5 days. shNT, non-specific shRNA. **e**, Levels of TGFβ-2 ligand in conditioned medium from A549 cells treated with 5 mM MMA ($n = 4$ independent experiments). **f**, **g**, *TGFB2* mRNA levels determined by qPCR (**f**; vehicle $n = 5$, MMA $n = 8$ mice) and immunoblots (**g**; representative images, $n = 8$ mice per group) in tumour samples from mice subcutaneously injected with the lower dose of MMA daily. **h**, Immunoblots of A549 cells treated with 5 mM MMA in the presence of TGFβ-neutralizing antibody; representative images ($n = 4$ independent experiments). **c**–**f**, Mean ± s.e.m., two-sided *t*-test. For gel source data, see Supplementary Fig. 2. ppSMAD3 S423/S425: SMAD3 phosphorylated on serine 423 and serine 425.

Having shown that MMA promotes pro-aggressive transcriptional remodelling through *SOX4* induction, we next sought to understand how MMA alters SOX4 levels. TCA-related metabolites are known for their ability to regulate transcription by modulating levels of histone methylation[29]. In the cancer cell models used here, however, treatment with MMA did not change total levels of major histone modifications (Extended Data Fig. 9a, b). Alternatively, the TGFβ pathway regulates SOX4 levels[30], and further analysis of the RNA sequencing (RNA-seq) data showed an increase in several components of the TGFβ signalling pathway, including the upregulation of TGFβ-2 ligand (Supplementary Table 2). We confirmed the increase in the mRNA of *TGFB2* (Extended Data Fig. 9c), which correlated with an increase over time in the abundance of TGFβ-2 in the medium of cancer cells treated with MMA (Fig. 3e). Moreover, and supporting the physiological relevance of these findings, analysis of tumour tissues from mice with elevated circulatory MMA levels showed a significant upregulation of *TGFB2* mRNA, a concomitant induction of TGFβ signalling, and upregulation of SOX4 (Fig. 3f, g). Time-course analysis showed that MMA robustly induced TGFβ signalling within 24 h, during which an increase in SOX4 was observed, before any of the pro-aggressive markers were detected (Extended Data Fig. 9d), suggesting a link between activation of TGFβ signalling, SOX4 induction and the acquisition of pro-aggressive properties driven by MMA. To test whether activation of TGFβ signalling is responsible for upregulation of SOX4, we concurrently treated cancer cells with MMA and a TGFβ receptor inhibitor or a pan-TGFβ neutralizing antibody. Both inhibition of the TGFβ receptor and neutralization of TGFβ ligand in the medium was sufficient to block the ability of MMA to induce SOX4 expression and pro-aggressive properties (Fig. 3h, Extended Data Fig. 9e). These data suggest that MMA relies on the activation of TGFβ signalling in an autocrine fashion to induce SOX4 and consequently the transcriptional reprogramming necessary for the cellular plasticity that sustains tumour progression.

Together, our results show that metabolic deregulation of the aged host plays a central role in the acquisition of aggressive properties that contribute to tumour progression. Specifically, ageing promotes an increase in circulatory MMA, which in turn endows cancer cells with the properties necessary to migrate, invade, survive and thrive as metastatic lesions, which results in decreased cancer-associated survival (Extended Data Fig. 9f). Although more in-depth studies are necessary to fully determine the scope of age-driven changes that contribute to the tumorigenic process, this study adds metabolic reprogramming to the complex relationship between ageing and cancer.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2630-0.

1. Wu, L. E., Gomes, A. P. & Sinclair, D. A. Geroncogenesis: metabolic changes during aging as a driver of tumorigenesis. *Cancer Cell* **25**, 12–19 (2014).
2. Thigpen, T. et al. Age as a prognostic factor in ovarian carcinoma. The Gynecologic Oncology Group experience. *Cancer* **71** (Suppl), 606–614 (1993).
3. Deng, F. et al. Age is associated with prognosis in serous ovarian carcinoma. *J. Ovarian Res.* **10**, 36 (2017).
4. Gloeckler Ries, L. A., Reichman, M. E., Lewis, D. R., Hankey, B. F. & Edwards, B. K. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist* **8**, 541–552 (2003).
5. White, M. C. et al. Age and cancer risk: a potentially modifiable relationship. *Am. J. Prev. Med.* **46** (Suppl 1), S7–S15 (2014).
6. Balducci, L. & Ershler, W. B. Cancer and ageing: a nexus at several levels. *Nat. Rev. Cancer* **5**, 655–662 (2005).
7. Serrano, M. & Blasco, M. A. Cancer and ageing: convergent and divergent mechanisms. *Nat. Rev. Mol. Cell Biol.* **8**, 715–722 (2007).

8. Migkou, M. et al. Short progression-free survival predicts for poor overall survival in older patients with multiple myeloma treated upfront with novel agent-based therapy. *Eur. J. Haematol.* **87**, 323–329 (2011).
9. Kasznicki, J., Sliwinska, A. & Drzewoski, J. Metformin in cancer prevention and therapy. *Ann. Transl. Med.* **2**, 57 (2014).
10. Li, J., Kim, S. G. & Blenis, J. Rapamycin: one drug, many effects. *Cell Metab.* **19**, 373–379 (2014).
11. De Lorenzo, M. S. et al. Caloric restriction reduces growth of mammary tumors and metastases. *Carcinogenesis* **32**, 1381–1387 (2011).
12. Ligibel, J. Lifestyle factors in cancer survivorship. *J. Clin. Oncol.* **30**, 3697–3704 (2012).
13. Ye, X. & Weinberg, R. A. Epithelial–mesenchymal plasticity: a central regulator of cancer progression. *Trends Cell Biol.* **25**, 675–686 (2015).
14. Wu, Y. & Zhou, B. P. Inflammation: a driving force speeds cancer metastasis. *Cell Cycle* **8**, 3267–3273 (2009).
15. Ferrucci, L. & Fabbri, E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat. Rev. Cardiol.* **15**, 505–522 (2018).
16. Zhu, Y., Carvey, P. M. & Ling, Z. Age-related changes in glutathione and glutathione-related enzymes in rat brain. *Brain Res.* **1090**, 35–44 (2006).
17. Tong, J. et al. Do glutathione levels decline in aging human brain? *Free Radic. Biol. Med.* **93**, 110–117 (2016).
18. Madeo, F., Carmona-Gutierrez, D., Kepp, O. & Kroemer, G. Spermidine delays aging in humans. *Aging (Albany NY)* **10**, 2209–2211 (2018).
19. Meynial-Denis, D. Glutamine metabolism in advanced age. *Nutr. Rev.* **74**, 225–236 (2016).
20. Chandler, R. J. & Venditti, C. P. Genetic and genomic systems to study methylmalonic acidemia. *Mol. Genet. Metab.* **86**, 34–43 (2005).
21. Riekeberg, E. & Powers, R. New frontiers in metabolomics: from measurement to insight. *F1000 Res.* **6**, 1148 (2017).
22. Baik, H. W. & Russell, R. M. Vitamin B12 deficiency in the elderly. *Annu. Rev. Nutr.* **19**, 357–377 (1999).
23. Zhang, J. et al. SOX4 induces epithelial-mesenchymal transition and contributes to breast cancer progression. *Cancer Res.* **72**, 4597–4608 (2012).
24. Wang, L. et al. SOX4 is associated with poor prognosis in prostate cancer and promotes epithelial-mesenchymal transition in vitro. *Prostate Cancer Prostatic Dis.* **16**, 301–307 (2013).
25. Yang, M. et al. Estrogen induces androgen-repressed SOX4 expression to promote progression of prostate cancer cells. *Prostate* **75**, 1363–1375 (2015).
26. Liao, Y. L. et al. Identification of SOX4 target genes using phylogenetic footprinting-based prediction from expression microarrays suggests that overexpression of SOX4 potentiates metastasis in hepatocellular carcinoma. *Oncogene* **27**, 5578–5589 (2008).
27. Lourenço, A. R. & Coffer, P. J. SOX4: joining the master regulators of epithelial-to-mesenchymal transition? *Trends Cancer* **3**, 571–582 (2017).
28. Vervoort, S. J. et al. Global transcriptional analysis identifies a novel role for SOX4 in tumor-induced angiogenesis. *eLife* **7**, e27706 (2018).
29. Wong, C. C., Qian, Y. & Yu, J. Interplay between epigenetics and metabolism in oncogenesis: mechanisms and therapeutic approaches. *Oncogene* **36**, 3359–3374 (2017).
30. Vervoort, S. J., van Boxtel, R. & Coffer, P. J. The role of SRY-related HMG box transcription factor 4 (SOX4) in tumorigenesis and metastasis: friend or foe? *Oncogene* **32**, 3397–3409 (2013).

Methods

Cell lines

MCF-10A human mammary epithelial cells were obtained from the American Type Culture Collection (ATCC), and were cultured as previously described[31]. HCC1806 human breast cancer (triple negative breast cancer; TNBC) and A549 human lung cancer (non-small cell lung cancer; NSCLC) cell lines were also obtained from ATCC and were cultured in RPMI 1640 medium (Corning) supplemented with 10% FBS (Sigma-Aldrich) and penicillin–streptomycin (Gibco). MDA-MB-231-luciferase cells as described previously[32] were generated from the MDA-MB-231 parental human breast cancer cell line (ATCC) in the Massague laboratory and obtained from Memorial Sloan Kettering Monoclonal Antibody Core facility. They were maintained in high-glucose DMEM (Gibco) supplemented with 10% FBS (Sigma-Aldrich) and penicillin–streptomycin (Gibco). HEK293T cells were obtained from GenHunter and cultured in high-glucose DMEM (Gibco) supplemented with 10% FBS (Sigma-Aldrich) and penicillin–streptomycin (Gibco). All cell lines were maintained at 37 °C and 5% CO_2. All cell lines were routinely tested for mycoplasma and were at all times mycoplasma-negative.

Mice

Female nu/nu athymic mice (Envigo) were purchased at the age of 4–6 weeks, and the experiments were started 7–10 days after the mice were received at the Weill Cornell Medicine Belfer Research Building Vivarium. Experimental groups of 7–10 mice were created randomly and mice were group housed (maximum five in a cage) in standard cages with unrestricted acidified water and food (PicoLab Rodent Diet 5053 (Labdiet, Purina) containing 20% protein and 5% fat). The only deviation from the standard housing was for animals that received MMA in their drinking water. Animal husbandry was carried out by the vivarium technical staff in a human xenograft designated area following animal biosafety level-2 procedures. The room was maintained at 21–23 °C with a 12-h light–dark cycle. The mice were maintained in compliance with Weill Cornell Medicine Institutional Animal Care and Use Committee protocols. The tumour size limit on the protocol was 20 mm on the largest dimension or 2.5 cm^3 tumour volume or 10% of body weight, whichever was reached first. For mouse studies no statistical method was used to predetermine sample size, mice were randomly distributed among the treatment groups and no blinding was performed.

Human serum

Human serum from 30 'young' (aged 30 and below) and 30 'old' (aged 60 and above) male adults with no diagnosed disease at the time of collection were obtained from BioreclamationIVT (now BioIVT) collected as two separate batches (15 young and 15 old donors in each batch). The authors of this paper did not participate in the recruitment of human participants or receive any patient identifiers. The vendor was responsible for recruitment of donors and sample collection. BioreclamationIVT collected sera from consented donors under their IRB-approved protocols at FDA-registered donor centres and their expansive clinical collection network. The specific serum used in this study is limited but samples from similar donors can be obtained from BioreclamationIVT. For detailed information on the donors please see Supplementary Table 1. In the figures, the race of donors is indicated as A, African American; C, Caucasian; H, Hispanic.

Cell culture treatments

To test the effects of aged serum in cancer cells, A549 or HCC1806 cells were seeded and their medium replaced the next day with medium containing 10% human serum for 4 days. Before replacement of the culture medium, the cells were washed three times with PBS. To evaluate the effects of the top upregulated metabolites in the serum of the elderly, A549 cells were plated in normal culture medium and treated with 5 mM quinolinate (QA) (Sigma-Aldrich), 5 mM phosphoenolpyruvate (PEP) (Sigma-Aldrich), 5 mM methylmalonic acid (MMA; Tocris), or vehicle (0.1% DMSO for quinolinate; double-distilled water for MMA and phosphoenolpyruvate) on the next day. For the subsequent treatments to determine the effects of MMA on cellular phenotypes, MCF-10A, HCC1806, MDA-MB-231-luciferase and A549 cells were seeded and treatment was initiated on the next day with indicated concentrations of MMA (Tocris) or vehicle (double-distilled water) for the time frames indicated. To test the specificity of MMA treatments in the phenotypes tested, A549 and HCC1806 cells were treated with 5 mM MMA (Tocris), 5 mM malonic acid (Sigma-Aldrich), 5 mM propionic acid (Sigma-Aldrich), 5 mM fumaric acid (Sigma-Aldrich), 5 mM pyruvic acid (Sigma-Aldrich), or vehicle (0.1% DMSO) for 10 days. MCF-10A cells were treated with vehicle (0.1% DMSO) or 1 mM each of the following acids: MMA (Tocris), malonic acid, propionic acid, succinic acid, pyruvic acid, hydroxyisobutyric acid (Sigma-Aldrich), fumaric acid (Sigma-Aldrich), maleic acid (Sigma-Aldrich), malic acid (Sigma-Aldrich), or α-ketoglutaric acid (Sigma-Aldrich) for 10 days. To test the effects of enhanced MMA permeability, A549 and HCC1806 cells were treated with 5 mM MMA or a range of dimethyl-MMA (Sigma-Aldrich) concentrations (5, 0.5, 0.05 or 0.005 mM), or vehicle (0.1% DMSO) for 10 days. MCF-10A cells, on the other hand, were treated with 1 mM MMA or a range of dimethyl-MMA concentrations (1, 0.1, 0.01 or 0.001 mM), or vehicle (0.1% DMSO) for 10 days. To test the role of TGFβ signalling in SOX4 induction upon MMA treatment, A549 cells were pretreated for 2 h with the TGFβ receptor inhibitor SB431542 (S1067; Selleck Chem) dissolved in DMSO or with 0.5 μg/ml TGFβ-neutralizing antibody (MAB1835, R&D Biosystems; normal mouse IgG from Santa Cruz sc2025 used as control), and then maintained with the inhibitor or the antibody for the duration of MMA treatment. For all acidic treatments, 25 mM HEPES (Sigma-Aldrich) was added to the treatment medium to buffer potential changes in pH, and the medium was replaced every day during the treatments.

Targeted metabolomics and data analysis

Circulatory polar metabolites were extracted using 80% (v/v) aqueous methanol as described before[33] from 100 μl sera from young and old donors. Targeted liquid chromatography-tandem mass spectrometry (LC–MS/MS) was performed using a 5500 QTRAP triple quadrupole mass spectrometer (AB/SCIEX) coupled to a Prominence UFLC HPLC system (Shimadzu) with Amide HILIC chromatography (Waters). Data were acquired in selected reaction monitoring (SRM) mode using positive/negative ion polarity switching for steady-state polar profiling of greater than 260 molecules. Peak areas from the total ion current for each metabolite SRM transition were integrated using MultiQuant v2.0 software (AB/SCIEX). Statistical analysis of the data was carried out using MetaboAnalyst v4.0, a free online software for the analysis of metabolomic experiments (https://www.metaboanalyst.ca/). The original data were normalized to the median of the entire metabolome in each sample and log-transformed before further analysis (Supplementary Table 1).

Proteomic analysis of HS

Abundant serum proteins were depleted using High Select Top 14 spin columns (Thermo #A36370) following the manufacturer's protocol. In brief, 10 μl serum was applied to each column and incubated for 10 min with end-over-end rotation. Depleted samples were collected by centrifugation at 2,000*g* for 2 min. Ice-cold 100% trichloroacetic acid was added to 20% final concentration. Proteins were allowed to precipitate on ice for 60 min and then pelleted for 10 min at 20,000*g* at 4 °C. Pellets were washed twice with ice-cold acetone and allowed to dry at room temperature. Dry protein pellets were re-suspended in 8 M urea, 50 mM ammonium bicarbonate (ambic). Proteins were reduced by addition of dithiothreitol (DTT) to 5 mM and incubation at room temperature for 30 min, then alkylated by adding iodoacetamide to 15 mM and incubating in the dark at room temperature for 30 min.

Iodoacetamide was quenched with an additional 5 mM DTT. Samples were diluted to 2 M urea with 50 mM ambic and digested overnight at room temperature by adding 600 ng lysyl endopeptidase (lysC, Wako Chemicals USA, Inc.). Samples were further diluted to 1 M urea with 50 mM ambic and digested with 600 ng sequencing grade modified trypsin (Promega) for 6 h at 37 °C with shaking. Digests were acidified by the addition of neat formic acid (FA) to 2% final concentration, and desalted on hand-packed C18 STAGE Tips[34]. Eluted peptides were dried in a centrifugal evaporator. Peptides were labelled with 10-plex amine reactive TMT labelling reagents (Thermo Fisher, Rockford, IL) by re-suspending in 100 µl of 0.2 M HEPES pH 8 and adding 0.2 mg of each TMT label in 10 µl anhydrous acetonitrile and incubating at room temperature for 1 h. Reactions were quenched with 8 µl 5% hydroxylamine and then acidified with 16 µl neat FA. Test mixtures for each 10plex set were generated by mixing 5 µl from each channel and analysed with a 75 min gradient version of the final analysis method (described below). The final mix was adjusted on the basis of this analysis to generate equal total reporter ion intensities from each TMT channel. Mixed peptides were desalted on 50 mg tC18 Sep-Pak cartridges (Waters, Milford, MA), dried, and re-suspended in 5 µl 5% FA. Mass spectrometric analysis was performed on a Thermo Orbitrap Fusion mass spectrometer (Thermo Fisher, Waltham, MA) equipped with an Easy nLC-1000 UHPLC (Thermo Fisher Scientific). Peptides were separated with a gradient of 6–25% ACN in 0.1% FA over 155 min and introduced into the mass spectrometer by nano-electrospray as they eluted off a self-packed 40 cm, 75 µm (ID) reverse-phase column packed with 1.8 µm, 120 Å pore size, C18 resin (Sepax Technologies, Newark, DE). They were detected using a data-dependent MS2 method with a real-time search (RTS) plugin[35] used to trigger MS3 scans for TMT reporter ion quantification. For each cycle, one full MS scan was acquired in the Orbitrap at a resolution of 120,000 with automatic gain control (AGC) target of 5×10^5 and a maximum ion accumulation time of 100 ms. Each full scan was followed by the selection of the most intense ions, as many as possible in 2 s total cycle time, for collision induced dissociation (CID) and MS2 analysis in the linear ion trap for peptide identification using an AGC target of 1.5×10^4 and a maximum ion accumulation time of 50 ms. Ions selected for MS2 analysis were excluded from reanalysis for 60 s. Ions with +1 or unassigned charge were also excluded from analysis. MS2 spectra were searched in real-time using the RTS module[35]. RTS settings required a binomial score threshold of 65 to trigger SPS MS3 scans using positively identified MS2 fragment ions. Selected MS2 ions were fragmented with a HCD collision energy of 55 and scanned in the orbitrap at a resolution of 50,000 at m/z 200. To increase coverage of lower-abundance proteins, we used the gene close-out feature to trigger a maximum of 10 MS3 per protein. MS/MS spectra were matched to peptide sequences using SEQUEST v.28 (rev. 13)[36] and a composite database containing the 20,415 Uniprot reviewed canonical predicted human protein sequences (http://uniprot.org, downloaded 1 May 2019) and its reversed complement. Search parameters allowed for three missed cleavages, a mass tolerance of 20 ppm, a static modification of 57.02146 Da (carboxyamidomethylation) on cysteine, and dynamic modifications of 15.99491 Da (oxidation) on methionine and 229.16293 for TMT on lysines and peptide amino termini. Peptide spectral matches (PSMs) were filtered to 1% FDR using the target-decoy strategy[37] combined with linear discriminant analysis (LDA)[38] using the SEQUEST Xcorr and ΔCn′ scores, precursor mass error, observed ion charge state, and the number of missed cleavages. The data were further filtered to a 1% protein FDR using the same strategy with protein scores derived from the product of all LDA peptide probabilities. Remaining peptide matches to the decoy database as well as contaminating proteins (for example, human keratins) were removed from the final data set. TMT reporter ion signal-to-noise (SN) values were extracted for all PSMs by identifying the maximum peak intensity within a 3-mD window around the theoretical m/z. Each PSM was required to have a sum reporter ion SN across all 10 TMT channels ≥ 100 for inclusion in subsequent protein quantification. Reporter ion intensities were adjusted to correct for the isotopic impurities of the different TMT reagents based on manufacturer supplied values. Protein quantification was performed separately for each 10plex by summing SN values from all matching PSMs for each channel. The protein sum SN values were normalized to correct for mixing errors by dividing each value by the sum of all values within its channel. These values were then transformed for each protein to generate a fractional intensity for each sample. All raw data files, peak lists, and the sequence database have been deposited in the MASSive repository (https://massive.ucsd.edu, ID: MSV000084974).

Measurements of MMA and vitamin B12 concentrations in HS

Frozen aliquots of the HS (unprocessed, delipidated, size excluded or lipidic structure depleted) were sent to ARUP Laboratories (Sat Lake City, Utah) for measurement of MMA (test code: 2005255) and vitamin B12 (test code: 0070150) concentrations. ARUP Laboratories is a national nonprofit and academic reference laboratory of diagnostic medicine.

Delipidation and size-exclusion in HS

HS samples were manipulated to assess the components of HS that might facilitate entrance of MMA into cells. To delipidate the HS, Cleanascite Lipid Removal Reagent (Biotech Support Group) was used according to the manufacturer's protocol specifically for serum samples, using a 1:4 volume ratio of Cleanascite reagent to sample. To deplete the serum from molecules larger than 3 kDa, serial filtration through size exclusion columns was performed. Initially serum was applied to Amicon Ultra-4 centrifugal filter units (Milipore) with a molecular size cut-off of 100 kDa and centrifuged at 4,000*g* at 4 °C. The flow-through fraction then was processed successively through filter units with molecular size cut-offs of 50, 10 and 3 kDa. Delipidated or size-excluded HS fractions were used in cell culture treatments similarly to unprocessed HS as described above.

Lipidic structure isolation from serum and depletion from HS

Lipidic structures were isolated from freshly thawed fetal bovine serum (FBS) or HS using Total Exosome Isolation (from Serum) reagent (Invitrogen) according to the manufacturer's protocol. In brief, 6 ml FBS or 1.5 ml of each HS was pelleted with the reagent (supernatants were saved to be used as LS-depleted serum) and then the pellets were resuspended in sterile PBS in a volume equal to half of the starting serum volume (3 ml and 750 µl, respectively). Aliquots of lipidic structures were kept at −80 °C to prevent freeze–thaw cycles. Cells were treated with 150 µl lipidic structures in PBS from HS (lipidic structures from equivalent to 300 µl HS) or 300 µl lipidic structure-depleted HS in 6 cm plates with 3 ml normal growth medium for a total of 4-day treatment. The medium was changed, and treatment repeated for a second time 48 h after the initial treatment. To control for presence of the Total Exosome Isolation reagent in these lipidic structure-depleted sera, we added the same amount of the reagent to cells treated with the control HS.

Complexing of MMA with artificial or serum-derived lipidic structures

MMA was complexed with artificial lipidic structures (Lipofectamine 2000) or with serum-derived lipidic structures to achieve final concentrations of 1, 10 or 100 µM in the culture medium. In brief, Lipofectamine 2000 and the indicated amounts of MMA were diluted in Opti-MEM (Gibco) medium and incubated for 30 min for complex formation. A549 cells were treated with the complexes and the medium was replaced with fresh growth medium 24 h later. Cells were treated one more time with freshly made complexes 48 h after the initial treatment. The medium was again replaced 24 h later and cell lysates were collected 4 days after the initial treatment with the complexes. For complexing MMA with lipidic structures isolated from either FBS or HS, the Exo-Fect Exosome Transfection Kit (System Biosciences) was used

according to the manufacturer's protocol. In brief, 200 µl lipidic structures from FBS or 300 µl lipidic structures from each HS was processed and resuspended in 600 µl PBS to be used to transfect 2–6 cm plates. Complexes were made fresh before the initial treatment and then the remaining complexes were stored at −80 °C until they were used for the second treatment. Cells were treated with 300 µl MMA-complexed lipidic structures in PBS in their normal growth medium that contained either FBS or horse serum. The cells were treated with the remaining 300 µl MMA-complexed lipidic structures a second time 48 h after the initial treatment. The cell lysates were collected 4 days after the initial treatment.

SOX4 gene silencing

sh*SOX4* no. 1 (TRCN0000018213), sh*SOX4* no. 2 (TRCN0000018214) and shNT (shGFP; TRCN0000072181, all from Sigma Aldrich) lentiviruses were produced by co-transfection of HEK293T cells with plasmids encoding psPAX2 (Addgene plasmid 12260), and pMD2.G (Addgene plasmid 12259) using X-tremeGENE HP (Roche) in accordance with the manufacturer's protocol. Medium was changed 24 h post-transfection and the virus harvested after 48 h, filtered, and used to infect MDA-MB-231 luciferase and A549 cells in the presence of 8 µg/ml polybrene (Sigma-Aldrich). Selection of resistant colonies was initiated 24 h later using 2 µg/ml puromycin (Sigma-Aldrich) for 24 h, after which the cells were treated with MMA as described above for the time frame indicated.

Quantification of TGFβ-2 ligand levels in cell culture medium

A549 cells were treated with 5 mM MMA for indicated amounts of time and then the conditioned culture medium was collected. The samples were centrifuged for 3 min at 300*g* to remove any cells and debris and activated using the Sample Activation Kit 1 (R&D Systems) according to the manufacturer's protocol. Human TGFβ-2 levels in these conditioned media were measured using the Human TGFB2 Quantikine ELISA Kit (R&D Systems) according to the manufacturer's protocol.

Chemotherapeutic drug assays

A549, HCC1806 and MCF-10A cells were treated with the indicated concentrations of MMA or vehicle (double-distilled water) for 10 days. A549 and MDA-MB-231-luciferase cells with *SOX4* silenced were also treated with MMA as described above, after which they were seeded in 96-well plates in technical triplicates. The cells were treated the next day with either vehicle control (DMSO (0.1%)), carboplatin (0–200 µM), or paclitaxel (0–7.5 nM) at various concentrations. The media containing the treatments were replaced every day for 4 days. At the end of the treatments the cells were fixed in 4% paraformaldehyde (Electron Microscopy Sciences) diluted in PBS for 30 min. After the fixative solution was removed, the plates were washed with PBS and stained with 0.1% crystal violet solution for 15 min. The staining solution was removed and the plates were washed three times under running water, to remove the excess stain, and allowed to dry at room temperature. To quantify the biomass, crystal violet staining was eluted with 100% methanol and the absorbance at 590 nm was measured using an Envision plate reader (Perkin Elmer).

Transwell migration and invasion assays

MCF-10A cells were trypsinized and collected as previously described[39]. Resuspension media were aspirated, and cells were resuspended in assay medium (DMEM/F12 (Corning), 0.5% Horse Serum (Gibco), 500 ng/ml hydrocortisone (Sigma), 100 ng/ml cholera toxin (Sigma)). For migration assays, Boyden chamber inserts (BD Biosciences, 8 µm pore size) were pre-coated with 25 µg/µl rat tail collagen 1 (Corning). Assay medium supplemented with 5 ng/ml EGF (Peprotech) was added to the bottom chamber of the cell culture inserts. Cells (5×10^4 cells per 250 µl assay medium) were then added to the top chamber of cell culture inserts in a 24-well companion plate. After 6 h of incubation, the cells that had migrated to the lower surface of the membrane were fixed with ethanol and stained with 0.2% crystal violet in 2% ethanol. For cell invasion assays, BD BioCoat invasion chambers coated with growth factor-reduced Matrigel were used. Invasion chambers were prepared according to the manufacturer's specifications and assays were performed as described for migration assays, except that 20 ng/ml of EGF (Peprotech) was added to MCF-10A assay medium to serve as the chemoattractant and cells were allowed to invade for 24 h.

For MDA-MB-231 and A549 cells, transwell migration and invasion assays were performed as described above with minor changes. For MDA-MB-231 cells, high-glucose DMEM (Gibco) supplemented with 250 µg/ml BSA (Sigma-Aldrich) was used as the assay medium, and high-glucose DMEM supplemented with 10% FBS (Sigma-Aldrich) was used as the chemoattractant for both migration (6 h) and invasion assays (20 h). For A549 cells, RPMI (Corning) supplemented with 250 µg/ml BSA (Sigma-Aldrich) was used as the assay medium, and RPMI medium supplemented with 10% FBS (Sigma-Aldrich) was used as the chemoattractant for the 24-h migration assay. Images of crystal violet-stained cells were captured using a Nikon DS-Fi2 camera, and quantifications were carried out in an automated way using Fiji/ImageJ v1.52. In brief, binary images of the area covered by crystal violet-positive cells was generated using thresholding and settings that were appropriate for control samples, and these settings were used throughout the analysis. The percentage area covered by crystal violet-positive cells was quantified for each condition, using a minimum of three technical replicates.

Analysis of CD24 and CD44

Cells were dissociated using Cell Stripper (Corning), collected on ice and pelleted by centrifugation. After removing the Cell Stripper and washing the cell pellet with ice cold PBS, the cells were stained on ice for 30 min in 100 µl DMEM/F12 (without phenol red) with an APC mouse anti-human CD44 (559942, BD Biosciences) and FITC mouse anti-human CD24 antibody (555427, BD Biosciences), or an APC mouse IgG2b (555745, BD Biosciences) and FITC mouse IgG2a (553456, BD Biosciences) as isotype controls. The antibodies were used at the dilution recommended by the manufacturer: 20 µl for 1×10^6 cells in a 100-µl test volume. After labelling, each sample was washed twice with ice-cold PBS and resolved on a BD Accuri C6 flow cytometer (BD Biosciences). Data were collected using the BD Accuri C6 flow cytometer software, and then data analysis to determine the medium fluorescence intensity (MFI) of CD24- and CD44-positive cells was performed using the FlowJo v10 software package. A representation of the gating strategy performed before MFI analysis can be seen in Supplementary Fig. 1.

Immunoblots for total cell lysates

Proteins were isolated directly from intact cells via acid extraction using a 10% TCA solution (10% trichloroacetic acid, 25 mM NH_4OAc, 1 mM EDTA, 10 mM Tris-HCl pH 8.0). Precipitated proteins were harvested and solubilized in a 0.1 M Tris-HCl pH 11 solution containing 3% SDS and boiled for 10–15 min. Protein content was determined with the DC Protein Assay kit II (BioRad), and 20 µg total protein from each sample was run on SDS–PAGE under reducing conditions. The separated proteins were electrophoretically transferred to a nitrocellulose membrane (GE Healthcare), which was blocked in TBS-based Odyssey Blocking buffer (LI-COR). Proteins of interest were probed with specific antibodies (listed as 'target protein' (catalog no. - vendor, dilution factor): E-cadherin (610181 - BD Biosciences, 1:1,000), ZO1 (5406S - Cell Signaling, 1:250), fibronectin (ab2413 - Abcam, 1:10,000), vimentin (5741S - Cell Signaling, 1:5,000), serpine1 (612024 - BD Biosciences, 1:1,000), CTGF (ab6992 - Abcam, 1:250), MMP2 (4022S - Cell Signaling, 1:500), SOX4 (ab80261 - Abcam, 1:100), SMAD3 (9523S - Cell Signaling, 1:500), ppSMAD3 S423/S425 (ab52903 - Abcam, 1:1,000), actin (sc1615 - Santa Cruz, 1:10,000). Membranes were incubated with the primary antibodies overnight at 4 °C, and then with the appropriate

horseradish peroxidase-conjugated (HRP) anti-rabbit (NA934 - GE Healthcare, 1:10,000), anti-mouse (NA931 - GE Healthcare, 1:10,000) or anti-goat (AP180P - Millipore, 1:10,000) immunoglobulin for 2 h at room temperature. The signals were developed using Amersham ECL detection system (GE Healthcare).

Analysis of histone post-translational modifications

A549 and MCF-10A cells were trypsinized and normalized for cell numbers. Cell pellets were washed twice with ice-cold PBS+ (PBS containing containing 5 mM sodium butyrate, 2 mM nicotinamide, 2 mM phenylmethylsulfonyl fluoride (PMSF), 10 mM Aprotinin, 10 mM Leupetin, 10 mM Pepstain A, 10 mM NaF, 10 mM $NaVO_4$, 0.02% (w/v) NaN_3) and then resuspended in Triton Extraction Buffer (TEB: PBS+ containing 0.5% Triton X 100 (v/v)) at a cell density of 10^7 cells per ml. Cell lysis was achieved on a rotator at 4 °C for 10 min. Nuclei were centrifuged at 6,500*g* for 10 min at 4 °C, and washed once with half the volume of TEB. Histones were extracted in 0.2 N HCl at a density of 4×10^7 nuclei per ml overnight at 4 °C. To remove debris, samples were centrifuged at 6,500*g* for 10 min at 4 °C. The histone proteins in the supernatant were collected and neutralized with 2 M NaOH at 1/10th of the volume of the supernatant. Eight micrograms of each sample was resolved on a 15% polyacrylamide gel and immunoblots were performed as described above. The following primary antibodies were used (listed as "target protein (catalog no. - vendor, dilution factor): H3K4me3 (61379 - Active Motif, 1:500), H3K27me3 (39155 - Active Motif, 1:500), H3K27ac (39133 - Active Motif, 1:500), H4K8ac (ab15823 - Abcam, 1:500), H3K56ac (39281 - Active Motif, 1:500), H3K9ac (ab4441 - Abcam, 1:500), H3K9me3 (ab8898 - Abcam, 1:1,000), H3K36me3 (ab9050 - Abcam, 1:500), Total H4 (ab10158 - Abcam, 1:1,000), Total H3 (4499s - Cell Signaling, 1:1,000).

Gene expression analysis

RNA was isolated using the PureLink RNA isolation kit (Life Technologies) and contaminant DNA was digested with DNase I (Amplification grade, Sigma-Aldrich). cDNA was synthesized using the iSCRIPT cDNA synthesis kit (BioRad) and analysed by qPCR using SYBR green master mix (Life Technologies) on a QuantStudio6 Real-Time PCR system with QuantStudio Real Time PCR software v1.3 (Life Technologies). Exported data were further processed in Microsoft Excel 2013. Target gene expression was normalized to expression of *TBP* and actin *ACTB*. Primer sequences can be found in Supplementary Table 3.

Global gene expression analysis (RNA-seq)

RNA from A549 cells treated with 5 mM MMA for 3 or 10 days was isolated as described above. Total RNA was sent to Active Motif for further processing and RNA-seq analysis. In brief, RNA quality was assessed using BioAnalyzer, and only RNAs with RIN values between 8.7 and 10.0 were used. RNA-seq libraries were prepared using the Illumina TruSeq RNA Sample Preparation v2 Guide (Illumina Part # 15026495). Polyadenylated RNA was enriched from 1 µg total RNA. Libraries were sequenced on Illumina NextSeq 500 as paired-end 42-nt reads, to a depth of 40.2–54.7 million read pairs. The TopHat algorithm v2.1.0 (Bowtie v2.2.6.0) was used to align the reads to the hg38 genome. which was obtained through iGenomes (https:/support.illumina.com/sequencing/sequencing_software/igenome.html). The alignments (37.5–51.4 million aligned pairs) in the BAM files were further analysed using the Cufflinks suite of programs v2.2.1 (running consecutively: Cufflinks v2.2.1.Linux_x86_64, Cuffcompare v2.2.1, Cuffdiff v2.2.2). Cufflinks was run using the hg38 genes as a reference database. The cufflinks outputs were compared using cuffdiff. The accession number for the raw sequencing data reported in this paper is GEO: GSE127001 and can be accessed on https://www.ncbi.nlm.nih.gov/geo/. A volcano plot visualizing the genes that were significantly changed by at least 1.5-fold was created using EnhancedVolcano v1.0.1[40]. The top 100 significantly changed genes were log-transformed and clustered using HierarchicalClustering on GenePattern v3.0[41] using Euclidean distance and pairwise centroid-linkage, and were row centred. Functional annotation analysis on genes that were significantly changed by at least 1.5-fold was performed using DAVID: Database for Annotation, Visualization, and Integrated Discovery v6.8[42]. Gene sets based on annotations from the Gene Ontology database (http://www.geneontology.org) were used. Only gene sets with 10 genes or more, and EASE score (a modified Fisher exact *P* value) of 0.05 or less were evaluated.

mRNA overlap analysis

The overlap between significantly altered mRNAs between A549 cells treated with MMA for 10 days (≥1.5-fold change) and significantly altered genes induced by SOX4 induction (≥2-fold change)[28] was evaluated using the GeneOverlap algorithm v.1.18.0[43]. Functional annotation analysis on the set of overlapping genes was performed using DAVID: Database for Annotation, Visualization, and Integrated Discovery[42]. Gene sets based on annotations from the Gene Ontology database (http://www.geneontology.org) were used. Gene sets with an EASE score of 0.05 or less were evaluated.

Lung colonization assay in mice

MDA-MB-231-lucifease cells were treated with human serum from young and old donors as described above before injection via the tail vein for evaluation of lung colonization. We used 11 young and 11 old HSs from the first batch, not including outliers. Each data point for each serum sample represents an average of data collected from three mice (used as technical replicates). To test the effect of MMA in lung colonization, luciferase-expressing MDA-MB-231 cells were treated with 1 or 5 mM MMA for 5 days. Similarly, *SOX4*-silenced MDA-MB-231-luciferase cells were also treated with 5 mM MMA for 5 days or with HS from old donors (six old HS from the first batch, not including the outliers, and six young HS combined with non-specific shRNA used as control; each data point for each serum sample represents an average of the data collected from three mice, used as technical replicates), after which they were injected into the tail vein and lung colonization was evaluated as described before[32,44]. In brief, female nu/nu athymic mice were injected with 100,000 cells in 100 µl PBS into the tail vein. Metastases were monitored using IVIS Spectrum CT Pre-Clinical In Vivo Imaging System (Perkin-Elmer). Each experimental group comprised 7–10 mice. After 6 weeks, luminescence was measured and quantified using Living Image Software v.4.5 (Perkin-Elmer) to determine lung colonization. All animal studies followed the guidelines of and were approved by the Weill Cornell Medicine Institutional Animal Care and Use Committee.

Orthotopic xenograft experiments in mice

To achieve increased circulatory MMA concentrations, female nu/nu athymic mice were treated as described before[45] with a dose escalation of a 'low' (100 µg MMA/g mouse followed by 200 µg MMA/g mouse) or 'high' (200 µg MMA/g mouse followed by 400 µg MMA/g mouse) dose, either by subcutaneous injections or through the drinking water. MMA treatments through either method did not cause behavioural changes or changes in drinking and feeding habits, nor did they cause changes in body weight throughout the duration of the experiments. In brief, mice were treated for 16 days with the lower dose of MMA for both concentration groups after which the dose was doubled until the end of the experiment. For subcutaneous injections, MMA (Sigma) was dissolved in 0.9% NaCl, adjusted to pH 7.4 with 6 M NaOH, and injections were performed daily during the time of the experiment. For delivery of MMA in the drinking water, MMA (Sigma) was dissolved in acidified water (as is standard IACUC procedure for mouse husbandry) and it was replaced with a fresh solution every 3.5 days. Tumours were established in the mice after the first 8 days of MMA treatment by injection of 2×10^6 MDA-MB-231-luciferase cells in 100 µl 50:50 PSB and Matrigel (Corning) into the third mammary fat pad on the right side of each mouse. Primary tumours and metastases were monitored using IVIS

Spectrum CT Pre-Clinical In Vivo Imaging System (Perkin-Elmer), and luminescence was measured and quantified using Living Image Software v.4.5 (Perkin-Elmer). To visualize metastatic spread, the primary tumours were covered and the upper body was imaged. The experiment continued until mice showed signs of significant illness or the primary tumours reached 10% of mouse weight, at which point they were killed as specified by IACUC. At the time of death, blood was collected to measure the serum MMA levels and primary tumour tissue was removed. Time of natural death or euthanasia was used to create the Kaplan–Meier curves. All animal studies followed the guidelines of and were approved by the Weill Cornell Medicine Institutional Animal Care and Use Committee. Frozen tumour tissue was powderized with a mortar and pestle over liquid nitrogen. RNA was isolated from the powdered tissue and the qPCR was performed as described in the *Gene expression analysis* section. For protein isolation powderized tissue was lysed with RIPA buffer for 30 min on ice and homogenized through a 22-gauge needle before being centrifuged at full speed for 10 min. The proteins present in the lysate supernatants were then quantified and processed as described under *Immunoblots for total cell lysates*.

Measurement of absolute MMA concentration

For intracellular measurements of MMA concentration, metabolites were extracted and processed as described in the targeted metabolomics section. For mouse sera, metabolite extraction was performed in a mixture of ice and dry ice as previously described[46,47]. In brief, 10 µl plasma was extracted with 800 µl 62.5% methanol containing 0.6 µg/ml glutaric acid, followed by an addition of 500 µl precooled chloroform. Samples were vortexed for 10 min at 4 °C and then centrifuged for other 10 min (maximum speed, 4 °C). After centrifugation, metabolites were separated into two phases divided by a protein layer: polar metabolites in the methanol/water (upper) phase and the lipid fraction in the chloroform (lower) phase. The samples were derivatized and measured as described before[48]. In brief, polar metabolites were derivatized for 90 min at 37 °C in 15 µl of 20 mg/ml methoxyamine in pyridine per sample. Subsequently, 15 µl *N*-(tert-butyldimethylsilyl)-*N*-methyl-trifluoroacetamide, with 1% tert-butyldimethylchlorosilane, was added to 7.5 µl of each derivative and incubated for 60 min at 60 °C. Polar metabolite fractions containing methylmalonic acid were dried at 4 °C in a vacuum concentrator. Methylmalonic acid concentrations were analysed by gas chromatography (7890A GC system) coupled to mass spectrometry (5975C Inert MS system) from Agilent Technologies. Metabolites were separated with a DB35MS column (30 m, 0.25 mm, 0.25 µm) using a carrier gas flow of helium fixed at 1 ml/min. A volume of 1 µl of sample was injected with a split ratio 1 to 3 with an inlet temperature set at 270 °C. For the detection of polar metabolites, the gradient was set at 100 °C for 1 min ramped to 105 °C at 2.5 °C/min, then to 240 °C at 3.5 °C/min and finally to 320 °C at 22 °C/min. For the measurement of metabolites by mass spectrometry, the temperatures of the quadrupole and the source were set at 150 °C and 230 °C, respectively. An electron impact ionization energy fixed at 70 eV was applied and scan mode was used for the measurement of polar metabolites ranging from 100 to 600 a.m.u. (mass). After acquisition by gas chromatography–mass spectrometry (GC–MS), an in-house Matlab (vR2016B) M-file was used to extract mass distribution vectors and integrated raw ion chromatograms. The natural isotopes distribution were also corrected using the method developed by Fernandez et al.[49]. Peak areas were normalized to those of the internal standard glutaric acid.

Statistical analysis

Data analyses were performed using Microsoft Excel 2013 and GraphPad Prism7. The two-tailed Student's *t*-test, two-way ANOVA and Mantel–Cox test were used to determine significance. In all types of statistical analysis values of $P < 0.05$ were considered significant. Data are represented as the mean ± s.e.m. (standard error of the mean) of individual data points of at least three independent samples. The number of independent samples and statistical method used in each experiment are reported in the figure legends. For all experiments similar variances between groups were observed. Normal distribution of samples was not determined. In the DAVID functional annotation analyses for the RNA-seq experiments the EASE score, a modified Fisher exact *P* value, was used to determine significance as recommended.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Information about the HS donors and the data from the metabolomics experiment can be found in Supplementary Table 1. RNA-seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (under accession code GSE127001) and are presented in Supplementary Table 2. All raw data files, peak lists, and the sequence database for the proteomics analysis have been deposited in the MASSive repository (https://massive.ucsd.edu) under ID MSV000084974. Other data supporting the findings of this study are available from the corresponding authors upon reasonable request. Source data are provided with this paper.

Code availability

The quantification of invasion/migration assay images were carried out in an automated way on Fiji/ImageJ v1.52 using a custom macro script. This macro is a basic automation script and cannot be used as standalone code, but it is available from the corresponding authors upon reasonable request.

31. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**, 256–268 (2003).
32. Minn, A. J. et al. Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518–524 (2005).
33. Yuan, M., Breitkopf, S. B., Yang, X. & Asara, J. M. A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat. Protoc.* **7**, 872–881 (2012).
34. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
35. Erickson, B. K. et al. Active instrument engagement combined with a real-time database search for improved performance of sample multiplexing workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).
36. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
37. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
38. Huttlin, E. L. et al. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
39. Shin, S., Dimitri, C. A., Yoon, S. O., Dowdle, W. & Blenis, J. ERK2 but not ERK1 induces epithelial-to-mesenchymal transformation via DEF motif-dependent signaling events. *Mol. Cell* **38**, 114–127 (2010).
40. Blighe, K. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. *R Package Version 1.0.1* (2019).
41. Reich, M. et al. GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
42. Dennis, G., Jr et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, 3 (2003).
43. Shen, L. & Sinai, M. GeneOverlap: Test and visualize gene overlaps. *R Package Version 1.18.0* (2018).
44. Oskarsson, T. et al. Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nat. Med.* **17**, 867–874 (2011).
45. Goyenechea, E. et al. Expression of proinflammatory factors in renal cortex induced by methylmalonic acid. *Ren. Fail.* **34**, 885–891 (2012).
46. van Gorsel, M., Elia, I. & Fendt, S.-M. in *Metabolic Signaling: Methods and Protocols Methods in Molecular Biology* (eds Lundt, Sophia & Fendt, S.-M.) chapter 4 (Humana, 2018).
47. Elia, I. et al. Proline metabolism supports metastasis formation and could be inhibited to selectively target metastasizing cancer cells. *Nat. Commun.* **8**, 15267 (2017).
48. Lorendeau, D. et al. Dual loss of succinate dehydrogenase (SDH) and complex I activity is necessary to recapitulate the metabolic phenotype of SDH mutant tumors. *Metab. Eng.* **43** (Pt. B), 187–197 (2017).

49. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of 13C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom*. **31**, 255–262 (1996).

Acknowledgements We thank members of the Blenis and Cantley laboratories for input; P. Coffer for the list of SOX4 targets; R. Pritchard for experimental assistance. A.P.G. is supported by a Susan G. Komen Postdoctoral Fellowship and a Pathway to Independence Award from NCI (K99CA218686). T.S. is supported by the NIH F31 pre-doctoral fellowship F31CA220750. J.F.-G. is supported by an FWO fellowship. C.K. was supported by a Medical Scientist Training Program grant from the NIGM/NIH T32GM007739 to the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program. This research was supported by NIH grant R01CA46595 to J.B. S.-M.F. is funded by the European Research Council under the ERC Consolidator Grant Agreement number 711486 – MetaRegulation, FWO research grants and projects, and KU Leuven Methusalem Co-funding.

Author contributions A.P.G. conceived the project. A.P.G. and D.I. performed all the molecular biology experiments, the EMT-related experiments and the invasion and migration experiments, prepared the RNA for RNA-seq experiments and assisted on all other experiments. A.P.M. quantified the migration and invasion experiments. T.S. evaluated the stemness markers. V.L., J.E.E. and D.B. performed the mouse experiments. A.R. performed the drug resistance assays, prepared the metabolites for metabolomic analysis, produced the viral particles and assisted in multiple other experiments. J.H. generated the constructs and the cell lines. C.K. and J.E.E. performed the assessment of histone post-translational modifications. D.B. and M.P. measured the concentration of MMA and analysed the data. A.A. and N.D. performed and analysed the results of the proteomic analysis of HS samples. T.S., E.M., I.E. and J.F.-G. assisted with the metabolomics experiments and helped with the metabolite treatments. J.A. performed the metabolomic analysis in HS. A.P.G., J.A., L.C.C., R.d.C., N.D., S.-M.F. and J.B. supervised the project. A.P.G., D.I., J.F.-G., M.P., D.B., T.S., V.L., J.E.E., and N.D. analysed the data. The manuscript was written by A.P.G. and edited by J.B., D.I., V.L., T.S., J.F.-G., I.E., D.B. and S.-M.F. All authors discussed the results and approved the manuscript.

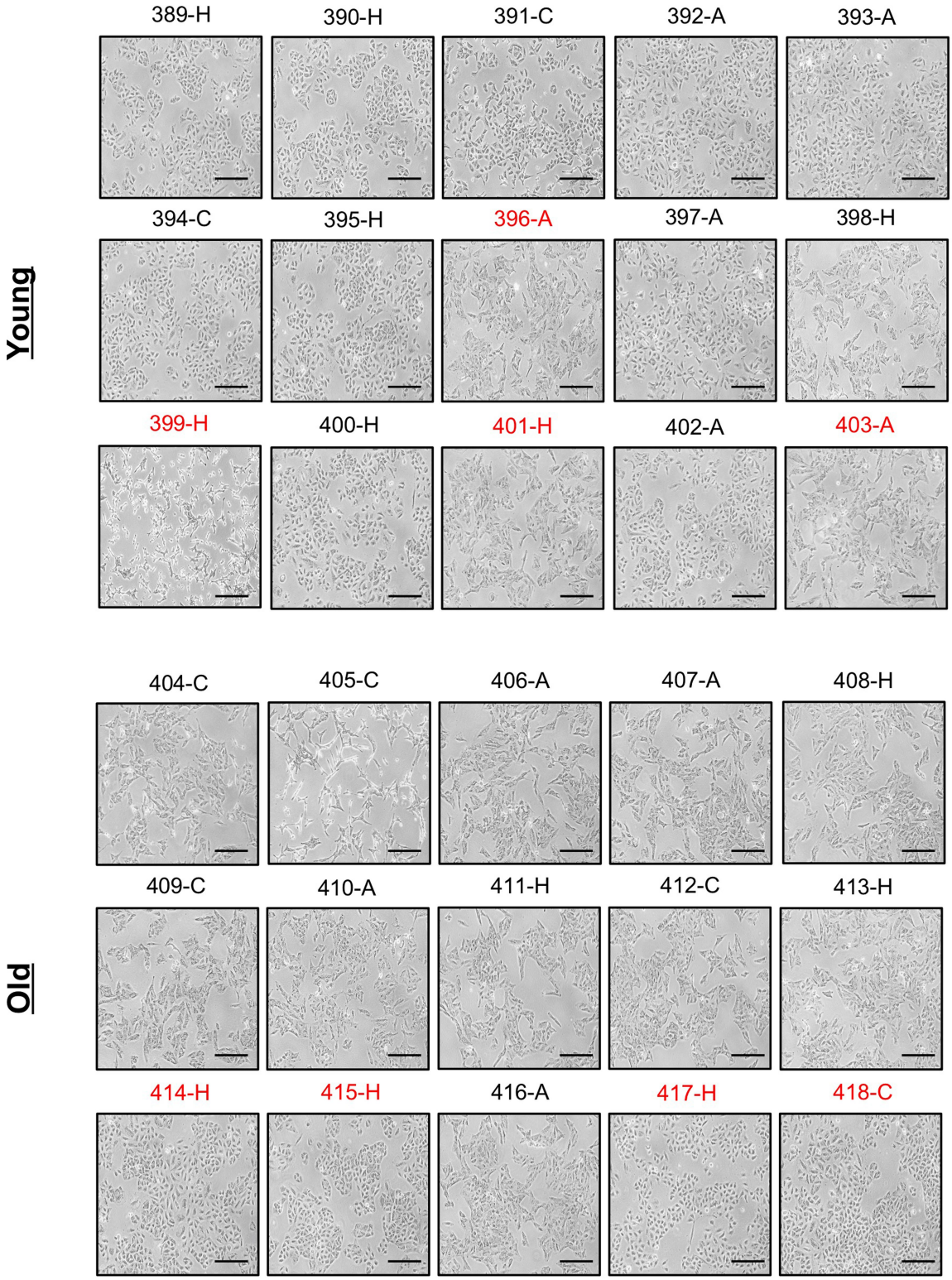
Competing interests L.C.C. owns equity in, receives compensation from, and serves on the Board of Directors and Scientific Advisory Board of Agios Pharmaceuticals and Petra Pharma Corporation. No potential competing interests were disclosed by the other authors.

Additional information

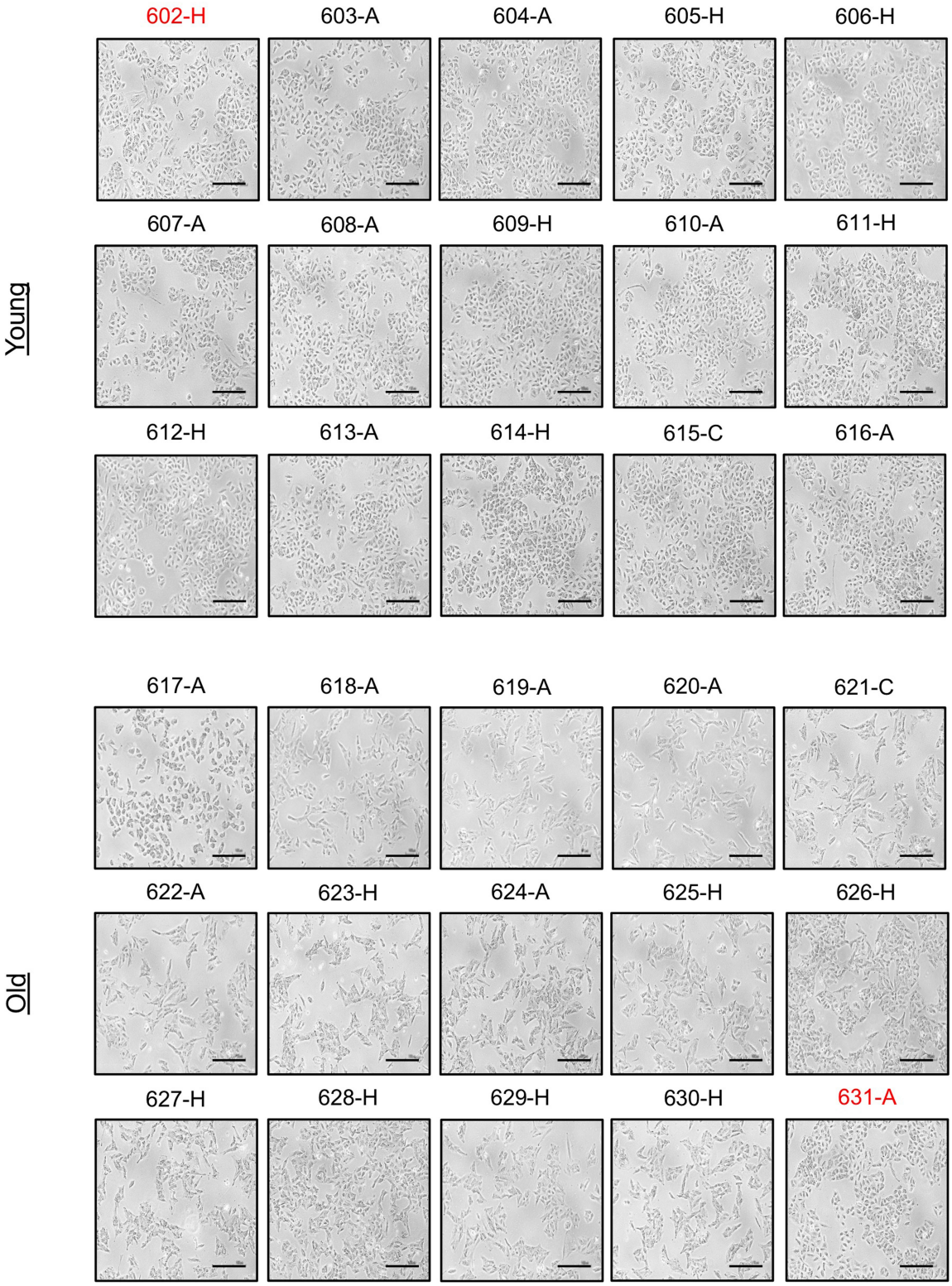
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2630-0.

Correspondence and requests for materials should be addressed to A.P.G. or J.B.

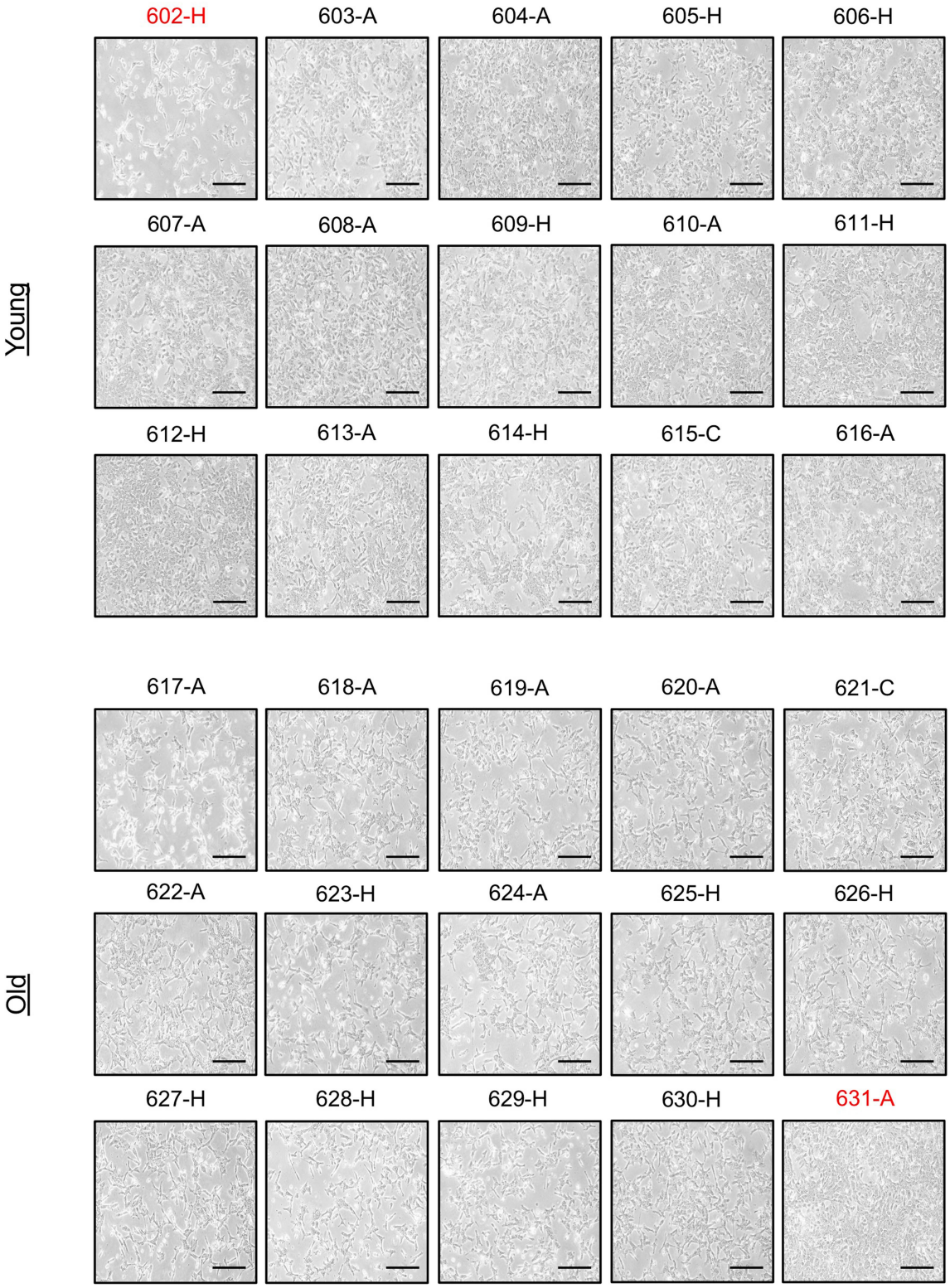
Peer review information *Nature* thanks Xiang H.-F. Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



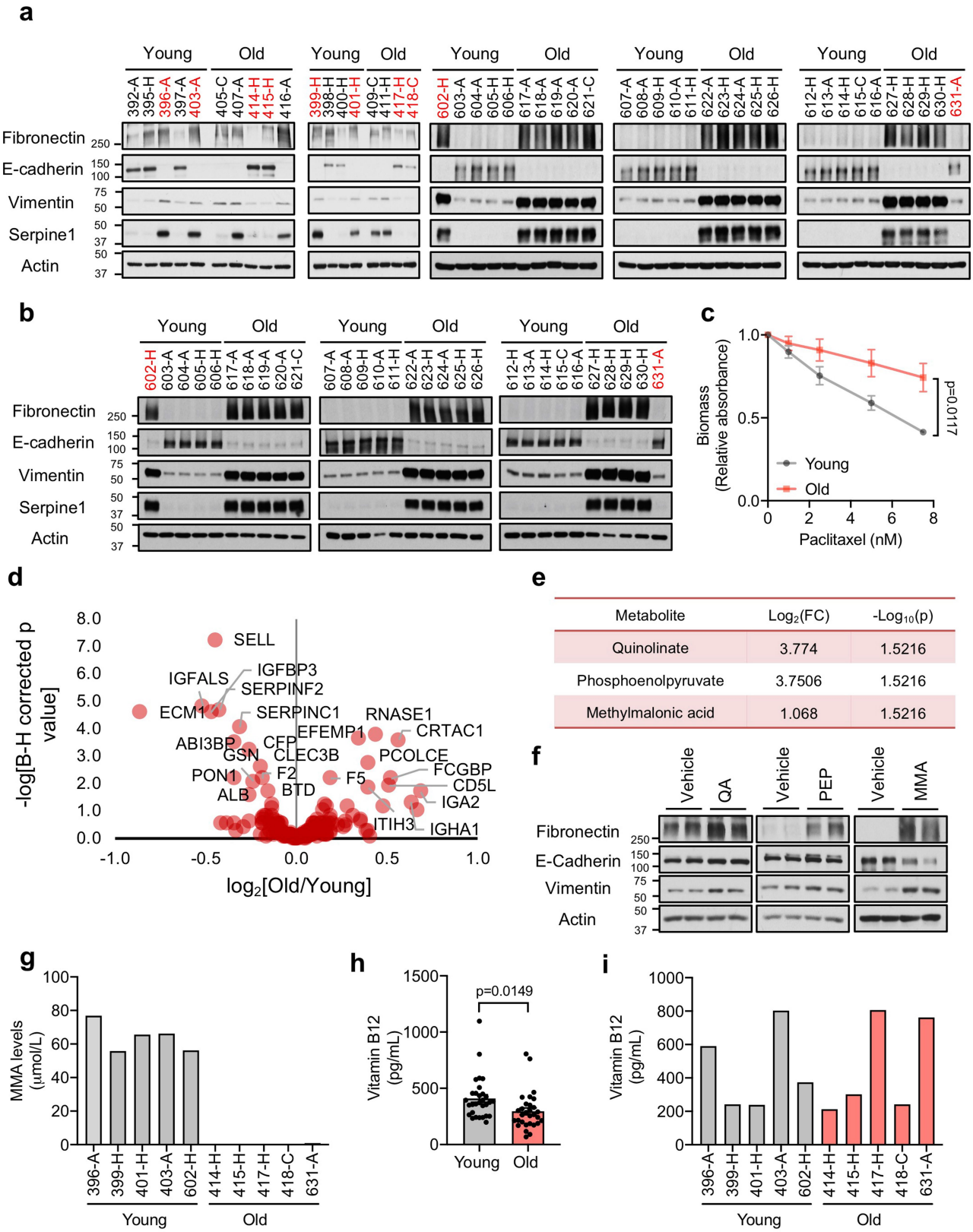
Extended Data Fig. 1 | Serum of old donors induces a mesenchymal-like phenotype in non-small cell lung cancer cells. Morphology of A549 cells cultured for 4 days with HS ($n = 15$ biologically independent samples per HS donor group, first batch of donors). Scale bar = 100 µm; red label indicates outlier donors.



Extended Data Fig. 2 | Serum of old donors induces an epithelial-to-mesenchymal transition phenotype in non-small cell lung cancer cells. Morphology of A549 cells cultured for 4 days with HS ($n = 15$ biologically independent samples per HS donor group, second batch of donors). Scale bar = 100 µm; red label indicates outlier donors.

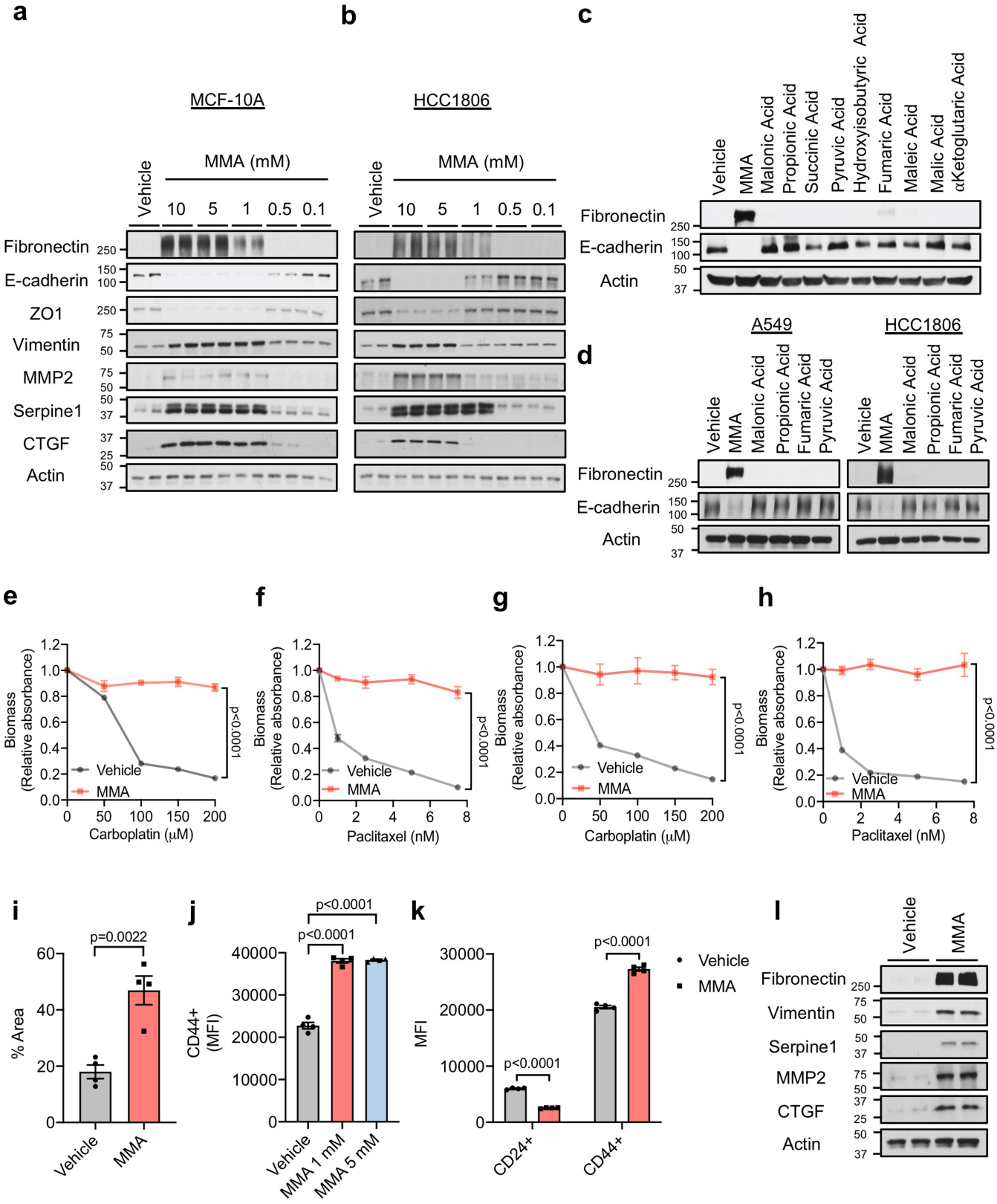


Extended Data Fig. 3 | Serum of old donors induces a mesenchymal-like phenotype in triple negative breast cancer cells. Morphology of HCC1806 cells cultured for 4 days with HS ($n = 15$ biologically independent samples per HS donor group, second batch of donors). Scale bar = 100 µm; red label indicates outlier donors.



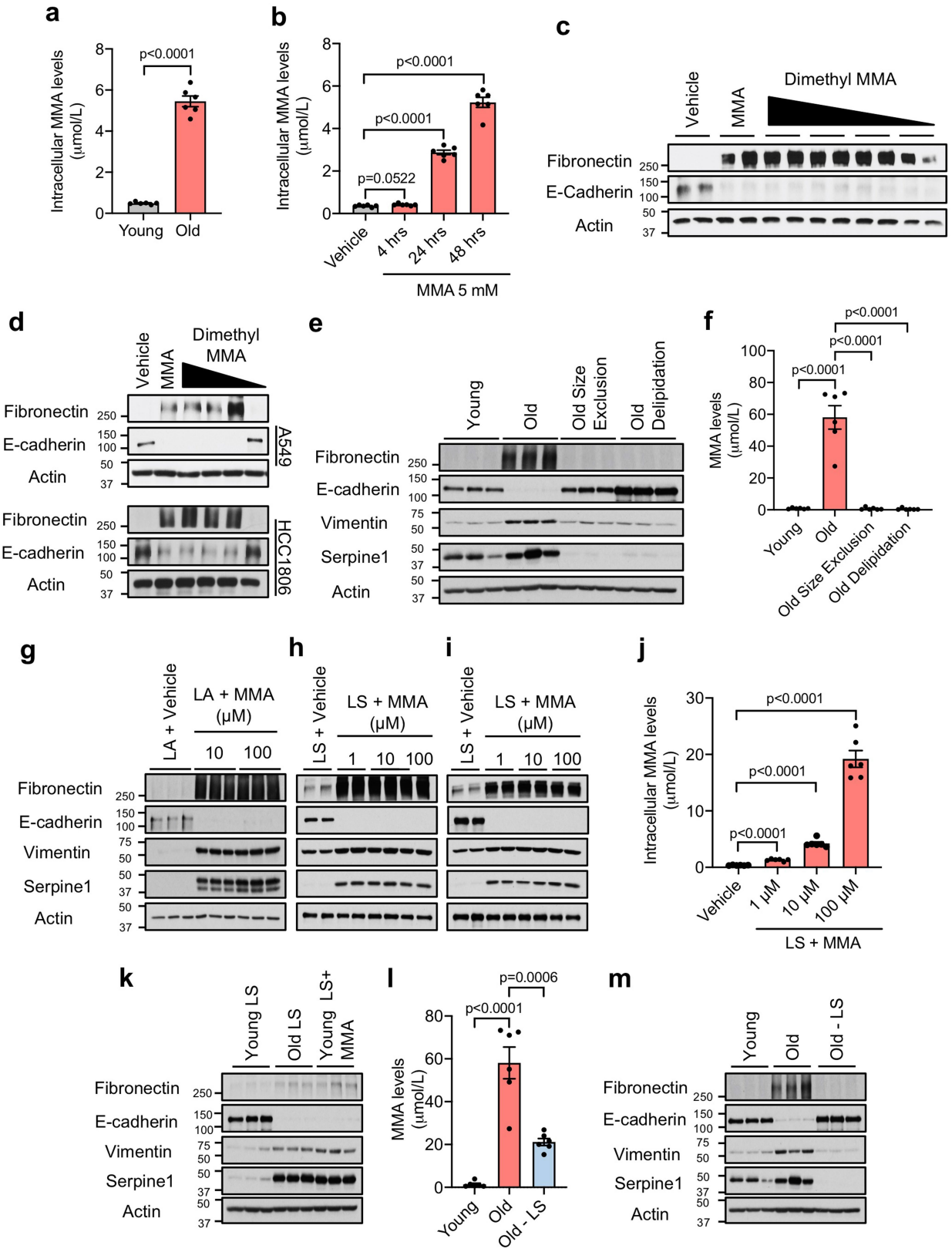
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Serum of old donors induces aggressive properties in cancer cells and displays a distinct metabolic profile. a, b, Immunoblots of A549 (total of $n = 30$ biologically independent samples per HS donor group, see Fig. 1b) (**a**) and HCC1806 ($n = 15$ biologically independent samples per HS donor group) (**b**) cells cultured for 4 days with HS **c**, Resistance to paclitaxel in A549 cells cultured for 4 days with HS ($n = 15$ biologically independent samples per HS donor group, two-way ANOVA). **d**, Volcano plot summarizing the proteomics analysis of all 60 human serum samples used in this study. **e**, List of all metabolites that are increased at a statistically significant level in the sera of old donors ($n = 11$ biologically independent HS donors, two-sided *t*-test). **f**, Immunoblots of A549 cells treated with 5 mM of each metabolite for 10 days; representative images ($n = 4$ independent experiments; QA: quinolinate, PEP: phosphoenolpyruvate, MMA: methylmalonic acid) **g**, Concentrations of MMA in 10 outlier human sera (serum from 5 young and 5 old donors, each bar represents the concentration of a single donor) **h, i**, Concentrations of vitamin B12 in all HS samples ($n = 30$ biologically independent samples per HS donor group, two sided *t*-test) (**h**) and 10 outlier human sera (serum from 5 young and 5 old donors, each bar represents the concentration of a single donor) (**i**). For (**c**, **h**) data are presented as mean ± SEM. For gel source data, see Supplementary Fig. 2.



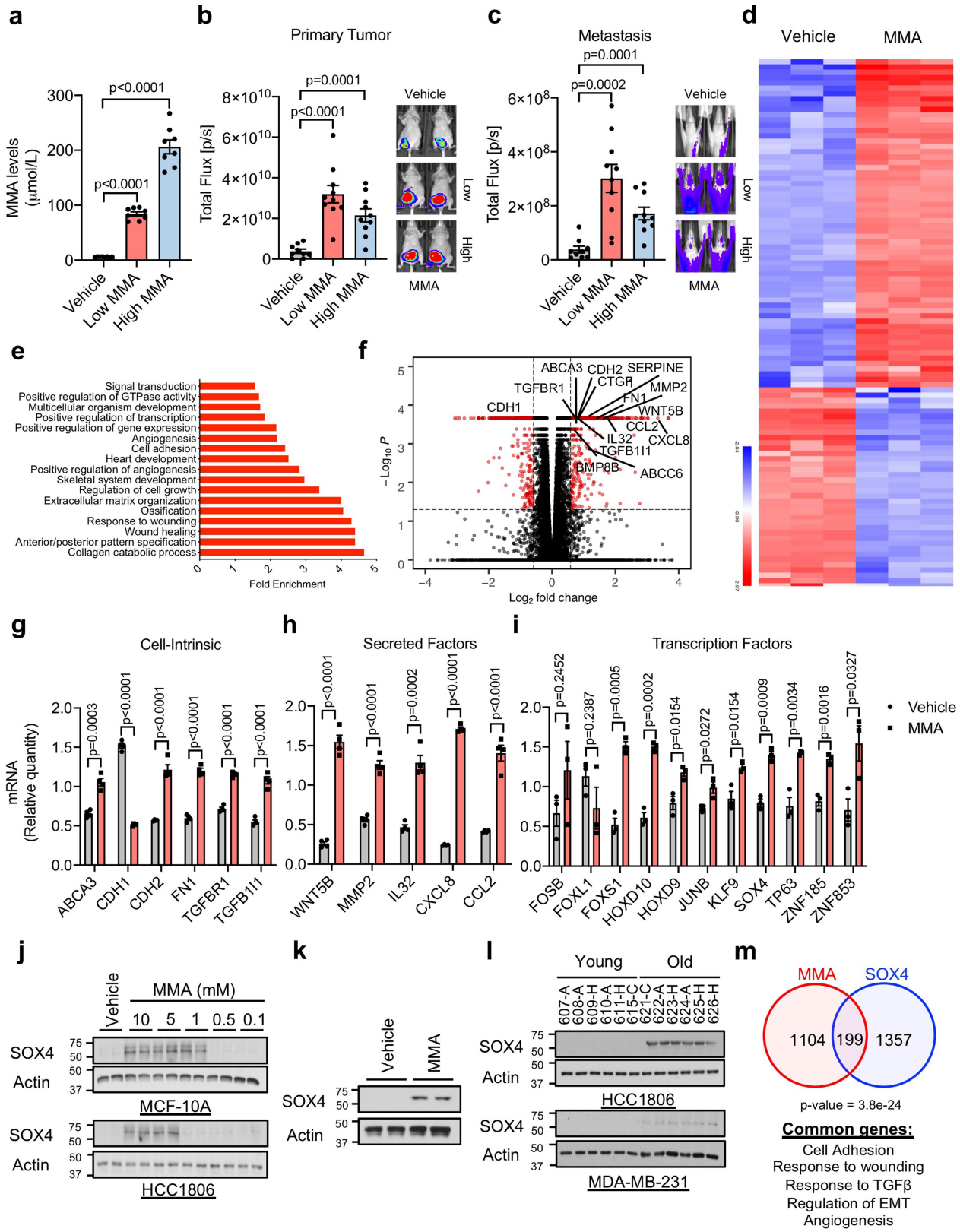
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Methylmalonic acid promotes epithelial-to-mesenchymal transition and metastatic-like properties. a, b, Immunoblots of MCF-10A (**a**) and HCC1806 (**b**) cells treated with MMA for 10 days ($n = 4$ independent experiments). **c, d**, Immunoblots of MCF-10A (**c**), A549 and HCC1806 (**d**) cells treated with 5 mM of the indicated acids for 10 days ($n = 4$ independent experiments). **e-h**, Resistance to carboplatin and paclitaxel in A549 (**e**, **f**) and HCC1806 (**g**, **h**) cells treated with 5 mM MMA for 10 days ($n = 4$ independent experiments, two-way ANOVA). **i**. Transwell migration assay of A549 cells treated with 5 mM MMA for 10 days ($n = 4$ independent experiments, two-sided *t*-test). **j, k**, Stemness evaluated by the increase in the CD44 marker in MCF-10A cells treated with MMA for 10 days (**j**) and by the increase in the CD44 marker and the decrease in CD24 marker in A549 cells treated with 5 mM MMA for 10 days (**k**) ($n = 4$ independent experiments, two-sided *t*-test). **l**, Immunoblots of MDA-MB-231-luciferase cells treated with 5 mM MMA for 5 days ($n = 4$ independent experiments). For (**e-k**) data are presented as mean ± SEM; (**a-d, l**) are representative images. For gel source data, see Supplementary Fig. 2.



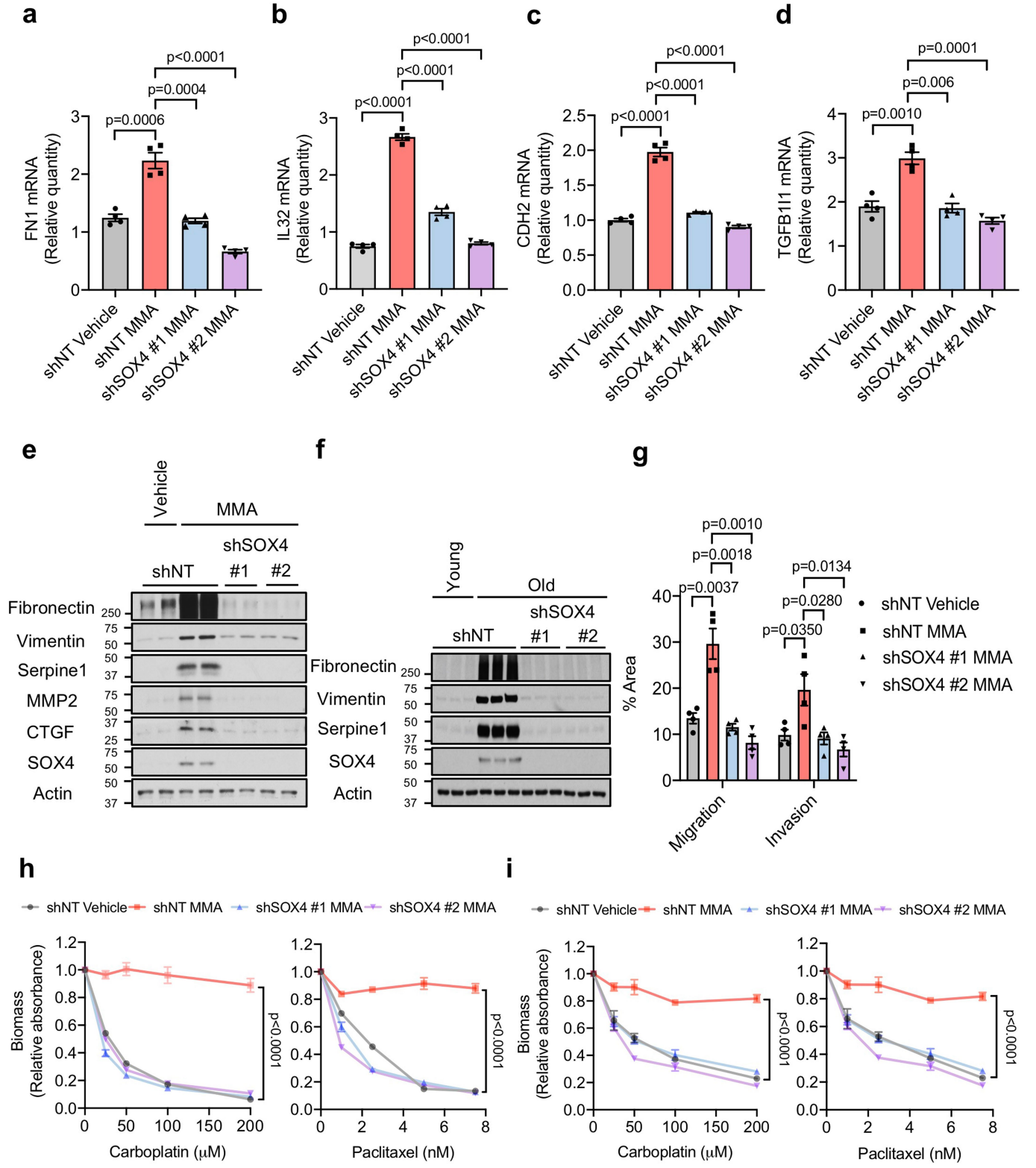
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Methylmalonic acid delivery is regulated by lipidic structures (LSs) in the sera of old donors. a, b, Intracellular MMA concentrations in A549 cells cultured with HS for 4 h ($n = 6$ biologically independent samples per HS donor group, two-sided t-test) (**a**) and with 5 mM MMA for indicated time periods ($n = 6$ independent experiments, two-sided t-test) (**b**). **c, d**, Immunoblots of MCF-10A (**c**), A549 and HCC1806 (**d**) cells treated with various concentrations (5–0.001 mM) of dimethyl methylmalonic acid for 10 days ($n = 4$ independent experiments). **e**, Immunoblots of A549 cells cultured for 4 days with young and old untreated HS, or old HS that was passed through size exclusion columns or delipidated ($n = 6$ biologically independent samples per HS donor group). **f**, MMA concentrations in the human serum used in (**e**) ($n = 6$ biologically independent samples per HS donor group, two-sided t-test). **g-i**, Immunoblots of A549 cells treated with complexes of lipofectamine (LA) with indicated amounts of MMA ($n = 3$ independent experiments) (**g**) and with LSs isolated from FBS that were complexed with indicated amounts of MMA ($n = 4$ independent experiments) (**h**), or of MCF10A treated with LSs isolated from FBS that were complexed with indicated amounts of MMA ($n = 4$ independent experiments) (**i**). **j**, Intracellular MMA concentrations in A549 cells cultured with MMA-loaded FBS lipidic structures (LSs) for 4 h ($n = 6$ independent experiments, two-sided t-test). **k**, Immunoblots of A549 cells treated with LSs isolated from young/old HS, and MMA-loaded LSs isolated from young HS (10 μM MMA) ($n = 6$ biologically independent samples per HS donor group). **l**, MMA concentrations in serum from old donors after depletion of LSs compared to control HS ($n = 6$ biologically independent samples per HS donor group, two-sided t-test). **m**, Immunoblots of A549 cells treated for 4 days with HS from old donors after depletion of LSs or from control donors ($n = 6$ biologically independent samples per HS donor group). For (**a, b, f, j, l**) data are presented as mean ± SEM; (**c, d, e, h, i, k, m**) are representative images. For gel source data, see Supplementary Fig. 2.

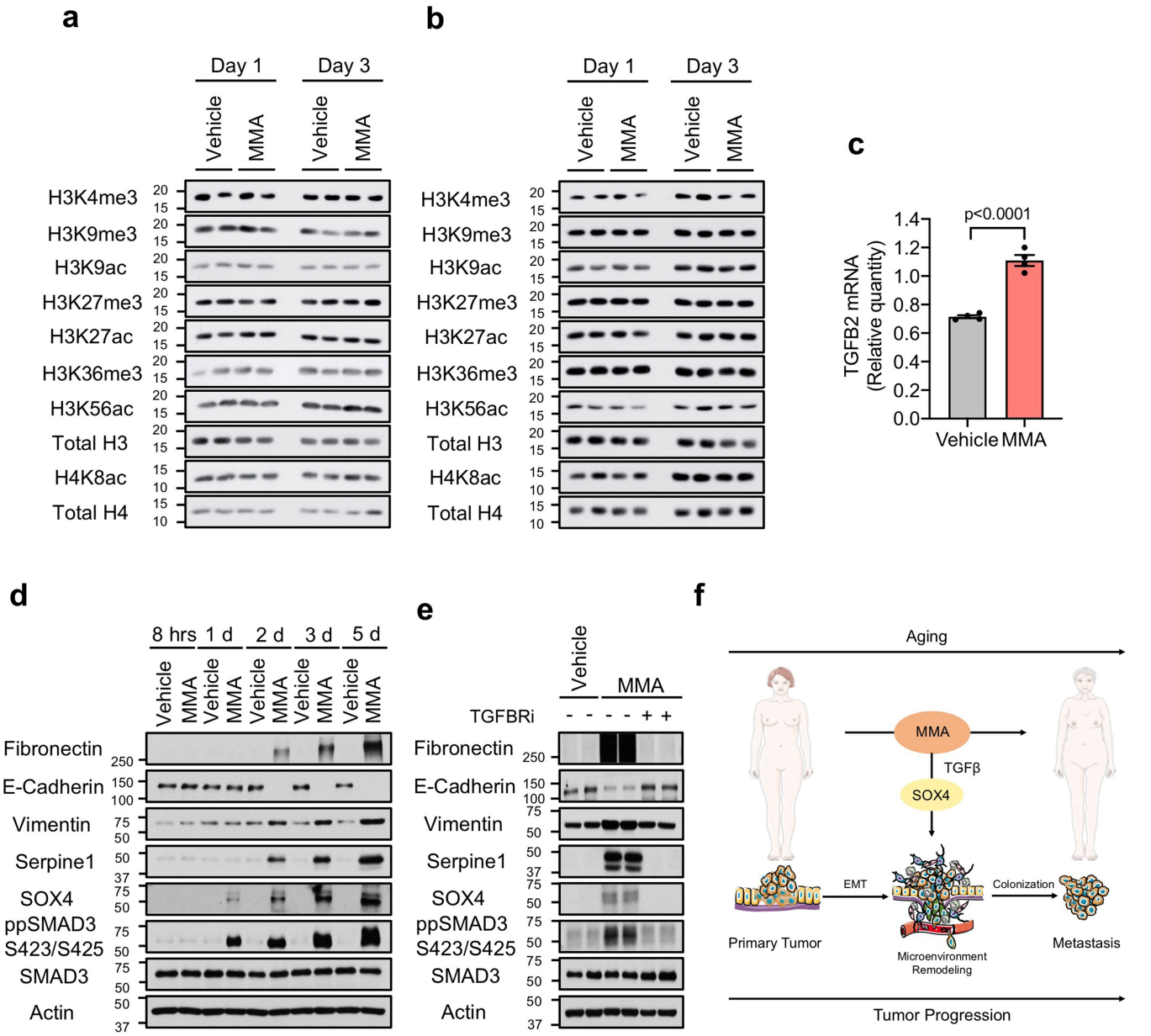


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Methylmalonic acid induces tumour progression through regulation of pro-aggressive and poor prognosis genes. a-c, End-point serum MMA concentrations ($n = 8$ mice per group, two-sided t-test) (**a**), bioluminescence intensity of the primary tumours ($n = 9$ mice on vehicle group and $n = 10$ on Low and High MMA group, two-sided t-test) (**b**), and metastases ($n = 10$ mice per group) (**c**) in mice that were xenografted with MDA-MB-231-luciferase cells and treated with MMA in their drinking water. **d-f**, Summary of RNA-seq analysis in A549 cells treated with 5 mM MMA for 10 days: a heatmap representation of hierarchical clustering of the top 100 changed mRNAs (**d**), functional annotation clustering analysis of the >1.5-fold changed mRNAs detected by RNA-seq analysis in A549 cells treated with 5 mM MMA for 10 days (**e**), and a volcano plot representation of the complete curated data set (the statistically significantly (FDR ≤ 0.05) altered mRNAs that are changed more than 1.5-fold are displayed in red) (**f**) ($n = 3$ independent experiments). **g**, **h**, mRNA levels of pro-aggressive cell intrinsic factors (**g**) and secreted factors (**h**) evaluated by qPCR in A549 cells treated with 5 mM MMA for 10 days ($n = 4$ independent experiments, two-sided t-test). **i**, mRNA levels of transcription factors evaluated by qPCR in A549 cells treated with 5 mM MMA for 3 days ($n = 3$ independent experiments, two-sided t-test). **j, k**, Immunoblots of MCF-10A and HCC1806 cells treated with MMA for 10 days (**j**) and of MDA-MB-231-luciferase cells treated with 5 mM MMA for 5 days (**k**) ($n = 4$ independent experiments) **l**, Immunoblots of HCC1806 and MD-MBA-231-luciferase cells cultured with HS for 4 days ($n = 6$ biologically independent samples per HS donor group) **m**, Venn diagram showing the overlap of altered mRNAs between A549 cells treated with MMA for 10 days and the genes altered by SOX4 induction[28] (Fisher's exact test). For (**a-c, g-i**) data are presented as mean ± SEM; (**j**, **k**) are representative images. For gel source data, see Supplementary Fig. 2.



Extended Data Fig. 8 | SOX4 mediates methylmalonic acid-induced pro-aggressive transcriptional reprogramming. a-d, mRNA levels of fibronectin (*FN*) (**a**), *IL32* (**b**), N-cadherin (*CDH2*) (**c**) and *TGFB1I1* (**d**) evaluated by qPCR in A549 cells with SOX4 knockdown and treated with 5 mM MMA for 10 days ($n = 4$ independent experiments, two-sided *t*-test). **e, f**, Immunoblots (**e**) and transwell migration/invasion assays (**f**) of MDA-MB-231-luciferase cells with SOX4 knockdown and treated with 5 mM MMA for 5 days ($n = 4$ independent experiments, two-sided *t*-test). **g**, Immunoblots of MDA-MB-231-luciferase cells with SOX4 knockdown and treated with HS for 5 days ($n = 6$ biologically independent samples per HS donor group) **h, i**, Resistance to carboplatin and paclitaxel in A549 cells with SOX4 knockdown and treated with 5 mM MMA for 10 days (**h**) and in MDA-MB-231-luciferase cells with SOX4 knockdown and treated with 5 mM MMA for 5 days (**i**) ($n = 4$ independent experiments, two-way ANOVA). For (**a-d, f, h, i**) data are presented as mean ± SEM; (**e, g**) are representative images. For gel source data, see Supplementary Fig. 2.



Extended Data Fig. 9 | Methylmalonic acid induces SOX4 through activation of TGFβ signalling. a, b, Immunoblots for histone marks in A549 (**a**) and MCF-10A (**b**) cells treated with 5 mM MMA for 1 or 3 days ($n = 4$ independent experiments). **c**, TGFB2 mRNA levels determined by qPCR in A549 cells treated with 5 mM MMA for 3 days ($n = 4$ independent experiments, two-sided t-test). **d, e**, Immuno-blots of A549 cells treated with 5 mM MMA for the indicated time points (**d**) or with 5 mM MMA in the presence of TGFBR inhibitor for 5 days (**e**) ($n = 4$ independent experiments). **f**, Age-induced accumulation of circulatory MMA induces SOX4 expression through the TGFβ pathway and elicits a transcriptional reprogramming that supports aggressiveness, promoting tumour progression and metastasis formation; This illustration was created using the Smart Servier Medical Art library (https://smart.servier.com/), which is licensed under a Creative Commons Attribution 3.0 Unported License. For (**c**) data are presented as mean ± SEM; (**a**, **b**, **d**, **e**) are representative images. For gel source data, see Supplementary Fig. 2.

nature research

Corresponding author(s): John Blenis

Last updated by author(s): Jul 5, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For targeted metabolomics peak areas from the total ion current for each metabolite SRM transition were integrated using MultiQuant v2.0 software (AB/SCIEX). For qPCR, QuantStudio Real Time PCR software v1.3 (Life Technologies) was utilized. For FACS BD Accuri C6 software v1.0.264.21 was utilized. For xenografted cells in mice luminescence was measured and quantified using the Living Image Software v4.5 (Perkin-Elmer).
Data analysis	All data analyses are explained in detail in the online methods. General data analyses were performed using Microsoft Excel 2013 and GraphPad Prism7. For targeted metabolomics statistical analysis of the data was carried out using MetaboAnalyst v4.0, a free online software for the analysis of metabolomic experiments (www.metaboanalyst.ca). For FACS analysis FlowJo v10 software package was used. For quantification of invasion/migrasion assays image quantifications were carried out in an automated way on Fiji/ImageJ v1.52 using a custom macro script. This macro is a basic automation script and cannot be used as a stand alone code but it is available from the corresponding authors upon reasonable requests. For RNA-seq analysis: The TopHat algorithm v2.1.0 (Bowtie v2.2.6.0) was used to align the reads to the hg38 genome, which was obtained through iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html). The alignments (37.5-51.4 M aligned pairs) in the BAM files were further analyzed using the Cufflinks suite of programs v2.2.1 (running consecutively: Cufflinks v2.2.1.Linux_x86_64, Cuffcompare v2.2.1, Cuffdiff v2.2.2). In addition the following suites were utilized: EnhancedVolcano v1.0.1 (Blighe, 2019), HierarchicalClustering on GenePattern v3.0 (Reich et al., 2006), DAVID: Database for Annotation, Visualization, and Integrated Discovery v6.8 (Dennis et al., 2003), the GeneOverlap algorithm v1.18.0 (Shen and Sinai, 2018). For proteomics analysis:MS/MS spectra were matched to peptide sequences using SEQUEST v.28 (rev. 13)6 and a composite database containing the 20,415 Uniprot reviewed canonical predicted human protein sequences (http://uniprot.org, downloaded 5/1/2019) and its reversed complement. Search parameters allowed for three missed cleavages, a mass tolerance of 20 ppm, a static modification of 57.02146 Da (carboxyamidomethylation) on cysteine, and dynamic modifications of 15.99491 Da (oxidation) on methionine and 229.16293 for TMT on lysines and peptide amino termini. Peptide spectral matches (PSMs) were filtered to 1% FDR using the target-decoy strategy7 combined with linear discriminant analysis (LDA)8 using the SEQUEST Xcorr and ΔCn' scores, precursor mass error, observed ion charge state, and the number of missed cleavages. The data were further filtered to a 1% protein FDR using the same strategy with protein scores derived from the product of all LDA peptide probabilities. Remaining peptide matches to the decoy database as well as contaminating proteins (e.g., human keratins) were removed from the final data set. TMT reporter ion signal-to-noise (SN)

values were extracted for all PSMs by identifying the maximum peak intensity within a 3 millidalton window around the theoretical m/z. Each PSM was required to have a sum reporter ion SN across all 10 TMT channels ≥ 100 for inclusion in subsequent protein quantification. Reporter ion intensities were adjusted to correct for the isotopic impurities of the different TMT reagents based on

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Public Databases Used: For proteomics analysis to match the peptide sequences a composite database was utilized containing the 20,415 Uniprot reviewed canonical predicted human protein sequences (http://uniprot.org, downloaded 5/1/2019) and its reversed complement. For RNA-seq analysis, the hg38 reference genome database was obtained through iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html) and the GSEA analysis was done with gene sets derived from the GO biological processes gene sets in the Molecular Signatures Database (MSigDB) collection v6.2.
Data Availability: RNA seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) under the accession code GSE127001, as well as in Supplementary Table 2. Source data information for the metabolomics experiment can be found on Supplementary Table 1. For the proteomics analysis all raw data files, peak lists, and the sequence database have been deposited in the MASSive repository (https://massive.ucsd.edu, ID#: MSV000084974). Source data and scans for the western blots are provided for all experiments. Other data that support the findings of this study are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were done. For in vivo studies, sample sizes were determined based on our previous experiments. In our experience, n = 7-10 mice per group is sufficient to detect meaningful biological differences with good reproducibility. For human serum treatment/analysis 30 young and 30 old serum was used from 3 different ethnicities, collected as two batches at different times to account for natural variability. For the mouse experiments and human sera treatment/analysis n refers to biologically independent samples as individual mice or human donors. The vast majority of the experiments outlined in this manuscript utilize cell-based systems, where conditions are standard and extreme variability is decreased as a result. Thus, we utilized at least 3 independent samples for each nominal variable. This number of independent experiments is a standard sample size to accurately detect differences in cell biology and molecular biology experiments (such as Napolitano et al. Nature 2020. https://doi.org/10.1038/s41586-020-2444-0.) Western blots are shown as representative images, however they were reproduced with at least 3 independent samples. Figure legends indicate when n refers to independent biological samples (such as when cells are treated with human sera) versus samples from independent experiments (such as when a cell line is treated with MMA).
Data exclusions	For the primary tumor and metastases luminescence data analysis, data from 1 mouse, which belonged to the vehicle drinking water group, were excluded since this mouse presented with abnormally large tumor and became very sick very quickly. Data on this mouse were still collected and excluded only after it was determined to be an outlier statistically determined by the Grubbs' Test to avoid investigator bias. For TGFB2 mRNA measurements from mice tumor tissues, 3 samples for the vehicle group were excluded. Due to limited supply of the tumor tissue from these 3 mice, RNA yield from these samples was insufficient to obtain qPCR data that pass quality standards.
Replication	Experiments were repeated multiple times with similar results as indicated in the figure legends. All attempts to replicate the data were successful. The core experiments were replicated by at least two people independently and in multiple cell lines to insure reproducibility and validity of the experiments. During mice experiments to achieve increased serum MMA levels MMA was introduced in the mice using two different ways (subcutaneous injections, and in drinking water) as a replicate.
Randomization	Mice obtained from the vendor were randomly split into experimental groups before administering xenografts. The donors for the human serum were selected by the vendor to only meet the age requirement we provided (young<30 and old>60), no other criteria or the purpose of the study was revealed to the vendor. No randomization was done for cell culture experiments, same plate of cells were used to set up the treatment groups in each experiment.
Blinding	For the human serum, the vendor did not know about the goal of the study as they selected the donors. Blinding was not performed for the in vitro and in vivo experiments. The investigators needed to know the treatment groups in order to perform the study, and the data analyses were based on objectively measurable data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Antibodies are listed as: target protein (catalog # - vendor, dilution used)
For immunoblots: E-Cadherin (610181 - BD Biosciences, 1:1,000), ZO1 (5406S - Cell Signaling, 1:250), Fibronectin (ab2413 - Abcam, 1:10,000), Vimentin (5741S - Cell Signaling, 1:5,000), Serpine1 (612024 - BD Biosciences, 1:1,000), CTGF (ab6992 - Abcam, 1:250), MMP2 (4022S - Cell Signaling, 1:500), SOX4 (ab80261 - Abcam, 1:100), SMAD3 (9523S - Cell Signaling, 1:500), ppSMAD3 S423/S425 (ab52903 - Abcam, 1:1,000), Actin (sc1615 - Santa Cruz, 1:10,000), H3K4me3 (61379 - Active Motif, 1:500), H3K27me3 (39155 - Active Motif, 1:500), H3K27ac (39133 - Active Motif, 1:500), H4K8ac (ab15823 - Abcam, 1:500), H3K56ac (39281 - Active Motif, 1:500), H3K9ac (ab4441 - Abcam, 1:500), H3K9me3 (ab8898 - Abcam, 1:1,000), H3K36me3 (ab9050 - Abcam, 1:5001:500), Total H4 (ab10158 - Abcam, 1:1,000), Total H3 (4499s - Cell Signaling, 1:1,000); HRP-conjugated anti-rabbit (NA934 - GE Healthcare, 1:10,000), anti-mouse (NA931 - GE Healthcare, 1:10,000) or anti-goat (AP180P - Millipore, 1:10,000) secondaries.
For flow cytometry (antibodies were used at the dilution recommended by the manufacturer , i.e., 20 µl for 1 x 10^6 cells in a 100- µl test volume): APC mouse anti-human CD44 (559942, BD Biosciences) and FITC mouse anti-human CD24 (555427, BD Biosciences), or an APC mouse IgG2b (555745, BD Biosciences) and FITC mouse IgG2a (553456, BD Biosciences).
For cell culture treatment: 0.5 µg/ml TGFβ neutralizing antibody (MAB1835, R&D Biosystems; normal mouse IgG from Santa Cruz sc2025 used as control). Multiple lot numbers have been utilized.

Validation

All antibodies used are commercially available and were tested by the manufacturer. They are standard tools used in the field, and have been previously validated and characterized by multiple labs. More information about each antibody and links to product citations can be found on manufacturers' websites. The Sox4 antibody was further validated by means of knockdown and reported in this manuscript. For flow cytometry, isotype controls were used to accurately determine the thresholds.

E-Cadherin (610181 - BD Biosciences) : https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/purified-mouse-anti-e-cadherin-36e-cadherin/p/610181
ZO1 (5406S - Cell Signaling): https://www.cellsignal.com/products/primary-antibodies/zo-1-antibody/5406
Fibronectin (ab2413 - Abcam): https://www.abcam.com/fibronectin-antibody-ab2413.html
Vimentin (5741S - Cell Signaling): https://www.cellsignal.com/products/primary-antibodies/vimentin-d21h3-xp-rabbit-mab/5741
Serpine1 (612024 - BD Biosciences): https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/cell-biology-reagents/cell-biology-antibodies/purified-mouse-anti-pai-1-41pai-1/p/612024
CTGF (ab6992 - Abcam): https://www.abcam.com/ctgf-antibody-ab6992.html
MMP2 (4022S - Cell Signaling): https://www.cellsignal.com/products/primary-antibodies/mmp-2-antibody/4022
SOX4 (ab80261 - Abcam): https://www.abcam.com/sox4-antibody-ab80261.html
SMAD3 (9523S - Cell Signaling): https://www.cellsignal.com/products/primary-antibodies/smad3-c67h9-rabbit-mab/9523
ppSMAD3 S423/S425 (ab52903 - Abcam): https://www.abcam.com/smad3-phospho-s423--s425-antibody-ep823y-ab52903.html
Actin (sc1615, Santa Cruz): https://www.scbt.com/p/actin-antibody-c-11
H3K4me3 (61379 - Active Motif): https://www.activemotif.com/catalog/details/61379
H3K27me3 (39155 - Active Motif): https://www.activemotif.com/catalog/details/39155/histone-h3-trimethyl-lys27-antibody-pab
H3K27ac (39133 - Active Motif): https://www.activemotif.com/catalog/details/39133/histone-h3-acetyl-lys27-antibody-pab
H4K8ac (ab15823 - Abcam): https://www.abcam.com/histone-h4-acetyl-k8-antibody-chip-grade-ab15823.html
H3K56ac (39281 - Active Motif): https://www.activemotif.com/catalog/details/39281/histone-h3-acetyl-lys56-antibody-pab
H3K9ac (ab4441 - Abcam): https://www.abcam.com/histone-h3-acetyl-k9-antibody-chip-grade-ab4441.html
H3K9me3 (ab8898 - Abcam): https://www.abcam.com/histone-h3-tri-methyl-k9-antibody-chip-grade-ab8898.html
H3K36me3 (ab9050 - Abcam): https://www.abcam.com/histone-h3-tri-methyl-k36-antibody-chip-grade-ab9050.html
Total H4 (ab10158 - Abcam): https://www.abcam.com/histone-h4-antibody-chip-grade-ab10158.html
Total H3 (4499s - Cell Signaling): https://www.cellsignal.com/products/primary-antibodies/histone-h3-d1h2-xp-rabbit-mab/4499
APC mouse anti-human CD44 (559942, BD Biosciences): https://www.bdbiosciences.com/us/applications/research/t-cell-immunology/t-follicular-helper-tfh-cells/surface-markers/human/apc-mouse-anti-human-cd44-g44-26-also-known-as-c26/p/559942
FITC mouse anti-human CD24 (555427, BD Biosciences): https://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/fitc-mouse-anti-human-cd24-ml5/p/555427
TGFβ neutralizing antibody (MAB1835, R&D Biosystems): https://www.rndsystems.com/products/tgf-beta1-2-3-antibody-1d11_mab1835

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	MCF-10A, A549, HCC1806 were bought from ATCC. Hek293T was bought from GenHunter. MDA-MB-231-luciferase cells described previously (Minn et al., 2005) were generated from the MDA-MB-231 parental cell line (ATCC) in Dr. Massague's lab and obtained from Memorial Sloan Kettering Cancer Center Monoclonal Antibody Core facility.
Authentication	ATCC utilizes multiple methods - morphology, karyotyping, and PCR based approaches - to authenticate the cell lines it maintains. No further authentication was performed by the authors of this manuscript.
Mycoplasma contamination	All cell lines were routinely tested for mycoplasma and were at all times mycoplasma negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female nu/nu athymic mice (Envigo) were purchased at the age of 4-6 weeks, and the experiments were started 7-10 days after the mice were received at the Weill Cornell Medicine Belfer Research Building Vivarium. Experimental groups of 7-10 mice were created randomly and mice were group housed, maximum 5, in standard cages with unrestricted water and food, PicoLab Rodent Diet 5053 (Labdiet, Purina) containing 20% protein and 5% fat. Only deviation from the standard housing was for animals that received MMA in their drinking water as described in "Orthotopic Xenograft Experiments in Mice" section. Animal husbandry was carried out by the vivarium technical staff in a human xenograft designated area following animal biosafety level-2 procedures. The room was maintained at 21-23°C and a 12h light-dark cycle. The mice were maintained in compliance to Weill Cornell Medicine Institutional Animal Care and Use Committee protocols.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field collected samples were used in this study.
Ethics oversight	The mice were maintained at Weill Cornell Medicine in compliance to Weill Cornell Medicine Institutional Animal Care and Use Committee protocols.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human serum from 30 "young" (aged 30 and below) and 30 "old" (aged 60 and above) male individuals with no diagnosed disease at the time of collection were obtained from BioreclamationIVT (now BioIVT) collected as two separate batches (15 young and 15 old donors in each batch). The specific serum used in this study is limited but samples from similar donors can be obtained from BioreclamationIVT (now BioIVT).
Recruitment	Investigators of this manuscript did not participate in the recruitment of human participants or received any patient identifiers. The vendor was responsible for recruitment of donors and sample collection.
Ethics oversight	BioreclamationIVT collected sera from consented donors under their IRB approved protocols at FDA registered donor centers and expansive clinical collection network.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were dissociated using Cell Stripper (Corning), collected on ice and pelleted by centrifugation. After removing the cell stripper media and washing the cell pellet with ice cold PBS, the cells were stained on ice for 30 minutes in 100 µL DMEM/F12

	(without phenol red) with an APC mouse anti-human CD44 (559942, BD Biosciences) and FITC mouse anti-human CD24 (555427, BD Biosciences) or an APC mouse IgG2b (555745, BD Biosciences) and FITC mouse IgG2a (553456, BD Biosciences) as isotype control. After labeling each sample was washed twice with ice cold PBS and resolved on BD Accuri C6 (BD Biosciences).
Instrument	BD Accuri C6 (BD Biosciences).
Software	Flowjo version 10
Cell population abundance	No sorting was done
Gating strategy	Cells were gated on FSC-A versus SSC-A excluding low size particles and the outliers of very high size and granularity. After that fluorescence intensity was analyzed (no further gating was performed). See supplemental figure 1 for the gating strategy.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Mitochondrial ubiquinol oxidation is necessary for tumour growth

https://doi.org/10.1038/s41586-020-2475-6

Received: 16 December 2018

Accepted: 20 April 2020

Published online: 8 July 2020

Check for updates

Inmaculada Martínez-Reyes[1], Luzivette Robles Cardona[1], Hyewon Kong[1], Karthik Vasan[1], Gregory S. McElroy[1], Marie Werner[1], Hermon Kihshen[1], Colleen R. Reczek[1], Samuel E. Weinberg[1], Peng Gao[2], Elizabeth M. Steinert[1], Raul Piseaux[1], G. R. Scott Budinger[1] & Navdeep S. Chandel[1,3]✉

The mitochondrial electron transport chain (ETC) is necessary for tumour growth[1–6] and its inhibition has demonstrated anti-tumour efficacy in combination with targeted therapies[7–9]. Furthermore, human brain and lung tumours display robust glucose oxidation by mitochondria[10,11]. However, it is unclear why a functional ETC is necessary for tumour growth in vivo. ETC function is coupled to the generation of ATP—that is, oxidative phosphorylation and the production of metabolites by the tricarboxylic acid (TCA) cycle. Mitochondrial complexes I and II donate electrons to ubiquinone, resulting in the generation of ubiquinol and the regeneration of the NAD+ and FAD cofactors, and complex III oxidizes ubiquinol back to ubiquinone, which also serves as an electron acceptor for dihydroorotate dehydrogenase (DHODH)—an enzyme necessary for de novo pyrimidine synthesis. Here we show impaired tumour growth in cancer cells that lack mitochondrial complex III. This phenotype was rescued by ectopic expression of *Ciona intestinalis* alternative oxidase (AOX)[12], which also oxidizes ubiquinol to ubiquinone. Loss of mitochondrial complex I, II or DHODH diminished the tumour growth of AOX-expressing cancer cells deficient in mitochondrial complex III, which highlights the necessity of ubiquinone as an electron acceptor for tumour growth. Cancer cells that lack mitochondrial complex III but can regenerate NAD+ by expression of the NADH oxidase from *Lactobacillus brevis* (*Lb*NOX)[13] targeted to the mitochondria or cytosol were still unable to grow tumours. This suggests that regeneration of NAD+ is not sufficient to drive tumour growth in vivo. Collectively, our findings indicate that tumour growth requires the ETC to oxidize ubiquinol, which is essential to drive the oxidative TCA cycle and DHODH activity.

To genetically decipher the mechanism that underlies the necessity of the ETC for tumour growth, we used 143B osteosarcoma cells that are deficient in mitochondrial complex III. These cells contain a four-base-pair deletion of the cytochrome *b* gene (143B-*CYTB*-Δ; *CYTB* is also known as *MT-CYB*), which encodes an essential component of complex III. The loss of complex III function results in dysfunctional ETC, oxidative phosphorylation (OXPHOS), and DHODH activities (Extended Data Fig. 1a, b). These cells maintain their mitochondrial membrane potential by reversing mitochondrial complex V (ATP synthase) activity[14]. 143B-*CYTB*-Δ cells have a negligible oxygen consumption rate (OCR) (Fig. 1a) and OXPHOS (that is, coupled OCR) (Extended Data Fig. 1c). 143B-*CYTB*-Δ cells are auxotrophic for pyruvate and uridine in vitro[2]. They require pyruvate to maintain levels of aspartate, a key metabolite for tumour growth in vivo, by maintaining the NAD+/NADH ratio[15–18], and uridine to maintain pyrimidine synthesis through the salvage pathway in the absence of DHODH activity due to loss of complex III function[19]. In the absence of cell-permeable methyl pyruvate or uridine, 143B-*CYTB*-Δ cells are unable to maintain aspartate synthesis or proliferate (Extended Data Fig. 1d–f). As expected, the whole-cell NAD+/NADH ratio (which is the average of the cytosolic and mitochondrial pools) was significantly lower in 143B-*CYTB*-Δ cells in the absence of methyl pyruvate and uridine (Extended Data Fig. 1g). Regardless of the availability of pyruvate and uridine, 143B-*CYTB*-Δ cells display significant differences in metabolite levels compared with wild-type 143B cells (Extended Data Fig. 1h, i). Our previous studies have demonstrated that 143B-*CYTB*-Δ cells in vitro can sustain anchorage-independent growth in the presence of pyruvate and uridine through glutamine-dependent reductive carboxylation[2,20]. However, 143B-*CYTB*-Δ cells were unable to grow tumours in vivo, which highlights different growth phenotypes between the in vitro and in vivo environments (Fig. 1b, Extended Data Fig. 1j). To further confirm the necessity of complex III for tumour growth in immunocompetent mice, we used CRISPR–Cas9 gene editing

[1]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [2]Robert H. Lurie Cancer Center Metabolomics Core, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [3]Department of Biochemistry and Molecular Genetics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
✉e-mail: nav@northwestern.edu

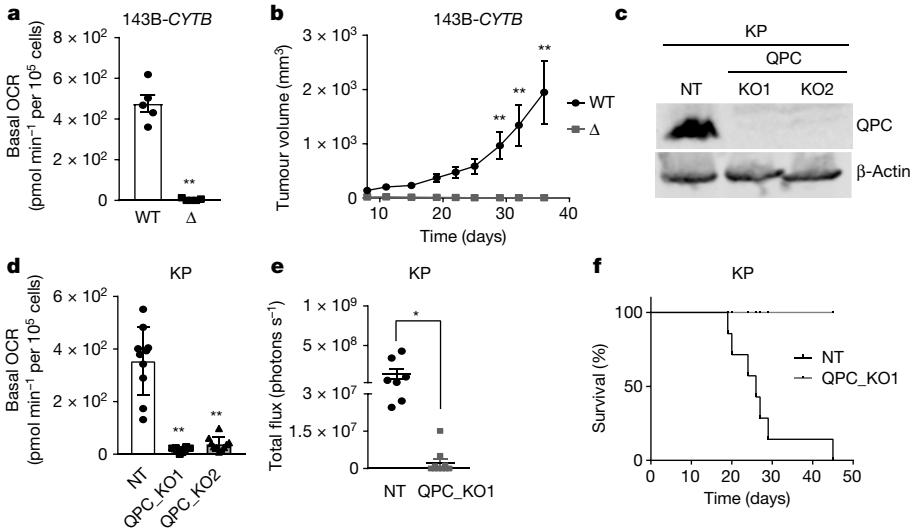


Fig. 1 | Complex III is necessary for tumour growth. a, Basal OCR of 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 5$ biologically independent experiments). **b**, Average tumour volume of xenografts from 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 10$ mice). **c**, Western blot analysis of QPC in KP-non targeting (NT) and knockout (KO) clones. β-actin was used as a loading control. **d**, Basal OCR of KP-NT and KP-QPC_KO cells ($n = 10$ replicates from two independent experiments). **e**, Luminescence values from the tumours. Values between days 19 and 33 after implantation with KP-NT cells ($n = 7$ mice), or day 33 after implantation with KP-QPC_KO cells ($n = 10$ mice). **f**, Survival of mice implanted with KP-NT ($n = 7$) and QPC_KO cells ($n = 10$). Data are mean ± s.e.m. (**a**, **b**, **e**) or mean ± s.d. (**d**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**a**, **e**), two-way analysis of variance (ANOVA) (**b**) with a Bonferroni test for multiple comparisons, or a one-way ANOVA (**d**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). Survival curves (**f**) were compared using the log-rank test ($P < 0.0001$). Tumour studies are from two independent cohorts. For gel source data, see Supplementary Fig. 1.

to knock out *Uqcrq*, which encodes the QPC subunit of complex III, in *Kras*$^{G12D/+}$*p53*$^{-/-}$ (KP; *p53* is also known as *Trp53*) cells isolated from mouse lung tumours (Fig. 1c). Loss of QPC in KP cells diminished basal and coupled OCRs (Fig. 1d, Extended Data Fig. 1k), and significantly reduced tumour growth after orthotopic mouse lung transplantation (Fig. 1e). Mice injected with non-targeting KP (KP-NT) cells had significantly worse survival than mice injected with QPC-knockout KP (KP-QPC_KO) cells (Fig. 1f). In addition, we explored the effect of the loss of complex III in T cell acute lymphoblastic leukaemia (T-ALL) in vivo (Extended Data Fig. 2a). Haematopoietic stem cells (HSCs) from donor mice with *loxP*-flanked (*Uqcrq*$^{flox/-}$) or wild-type (*Uqcrq*$^{WT/-}$) *Uqcrc* alleles and tamoxifen-inducible *Ubc-cre*^{ERT2} were transformed and adoptively transferred into immunocompetent mice. After establishment of T-ALL detectable in the peripheral blood of the recipients, tamoxifen was administered to induce the loss (QPC-KO) or maintenance (QPC-WT) of complex III function in T-ALL cells (Extended Data Fig. 2a). Analysis of GFP$^+$ T-ALL cell contents in the spleen and bone marrow revealed that only wild-type QPC cells were able to establish significant T-ALL burden (Extended Data Fig. 2b–e). Accordingly, the spleens of mice containing wild-type QPC T-ALL cells were significantly enlarged compared with those containing knockout QPC (Extended Data Fig. 2f), and mice containing leukaemic cells with functional mitochondria (wild-type QPC) had significantly worse survival (Extended Data Fig. 2g). Collectively, these data indicate that mitochondrial complex III is required for tumour growth in vivo.

Ubiquinol oxidation is an essential activity of mitochondrial complex III that allows complex I, II and DHODH to function. We ectopically and stably expressed GFP or *Ciona intestinalis* AOX in 143B-*CYTB*-Δ cells to restore ubiquinol oxidation[14] (Extended Data Fig. 3a). AOX transports electrons from ubiquinol directly to oxygen, bypassing ETC complex III and IV activities[12]. As a result, AOX restored the basal OCR in 143B-*CYTB*-Δ cells (Fig. 2a). AOX conducts electron flux but not proton pumping, thus it does not directly contribute to the proton-motive force for ATP synthesis. However, ubiquinol oxidation by AOX allows complex I to proton pump, consequently restoring OXPHOS (Extended Data Fig. 3b). AOX expression in 143B-*CYTB*-Δ cells alleviated their auxotrophy for pyruvate and uridine (Fig. 2b), restored the NAD+/NADH ratio (Fig. 2c), aspartate levels (Fig. 2d), and partially rescued TCA cycle metabolite levels in the absence of methyl pyruvate and uridine (Extended Data Fig. 1h). Notably, AOX expression in 143B-*CYTB*-Δ cells rescued tumour growth in vivo (Fig. 2e, Extended Data Fig. 3c). Similarly, AOX expression in KP-QPC_KO cells rescued basal and coupled OCR (Fig. 2f, Extended Data Fig. 3d), and in vivo lung tumour growth (Fig. 2g). Mice transplanted with KP-QPC_KO AOX-expressing cells had significantly worse survival than mice transplanted with KP-QPC_KO GFP-expressing control cells (Fig. 2h). Our results indicate that the essential function of mitochondrial complex III for tumour growth is ubiquinol oxidation and not its ability to proton pump or donate electrons to the downstream electron carrier cytochrome *c*.

Ubiquinol oxidation supports DHODH function (Extended Data Fig. 3a). Similar to the genetic inactivation of cytochrome *b*, treatment with the complex III inhibitor antimycin A rendered 143B-*CYTB*-WT cells auxotrophic for pyruvate and uridine (Extended Data Fig. 3e). However, the complex I inhibitor piericidin A made the cells auxotrophic for pyruvate but not uridine (Extended Data Fig. 3e). Notably, the dihydroorate-to-orotate ratio increased in 143B-*CYTB*-WT cells treated with antimycin A, but not with piericidin A (Extended Data Fig. 3f). These results indicate that the availability of the ubiquinone (Q) pool, which is only compromised when complex III function is inhibited, is the key factor for the maintenance of de novo pyrimidine synthesis. We tested the necessity of the de novo pyrimidine synthesis pathway through CRISPR–Cas9-mediated inactivation of *DHODH* in AOX-expressing 143B-*CYTB*-Δ cells (Extended Data Fig. 4a). Loss of DHODH caused uridine auxotrophy and reduced tumour growth in vivo (Extended Data Fig. 4b–e), and reconstituting its cDNA restored those phenotypes (Extended Data Fig. 4f–i).

Ubiquinol oxidation is required for mitochondrial complex I function. Therefore, we tested the necessity of complex I in AOX-expressing 143B-*CYTB*-Δ cells by inactivating *NDUFS2*, which encodes an essential subunit of complex I (Fig. 3a, Extended Data Fig. 5a). The loss of NDUFS2 made AOX-expressing 143B-*CYTB*-Δ cells auxotrophic for pyruvate in vitro (Fig. 3b), and ablated their in vivo tumour growth (Fig. 3c,

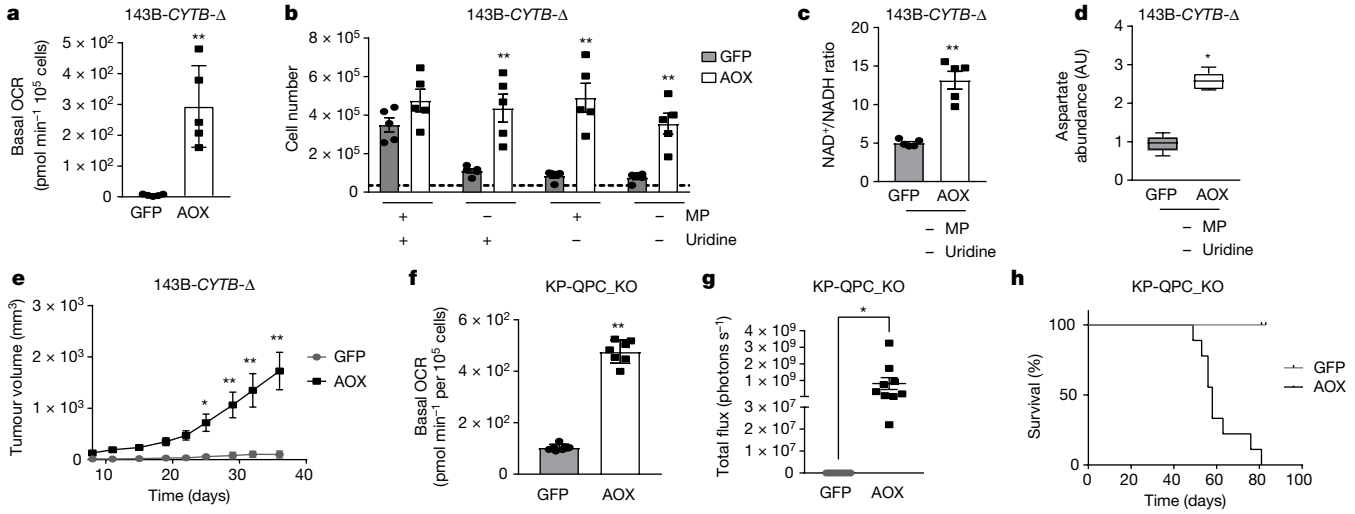


Fig. 2 | Ubiquinol oxidation by complex III is necessary for tumour growth. **a**, Basal OCR of 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells ($n = 5$ biologically independent experiments). **b**, 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells were grown in the presence or absence of methyl pyruvate (MP) and/or uridine, and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **c**, **d**, Intracellular NAD+/NADH ratio (**c**) and aspartate levels (**d**) of 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells in the absence of methyl pyruvate and uridine ($n = 5$ biologically independent experiments). **e**, Average tumour volume of xenografts from 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells ($n = 9$ mice). **f**, Basal OCR of KP-QPC_KO-GFP and KP-QPC_KO-AOX cells ($n = 7$ technical replicates; representative of five biologically independent experiments). **g**, Luminescence values from the tumours. Values before euthanasia between days 49 and 83 after implantation with KP-QPC_KO-AOX, or day 81 or 83 after implantation with KP-QPC_KO-GFP cells ($n = 9$ mice). **h**, Survival of mice implanted with KP-QPC_KO-GFP and KP-QPC_KO-AOX cells. ($n = 9$ mice). Data are mean ± s.e.m. (**a**–**e**, **g**) or mean ± s.d. (**f**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**a**, **c**, **f**, **g**) or two-way ANOVA (**b**, **e**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). Survival curves were compared using the log-rank test ($P < 0.0001$). Aspartate levels (**d**) were analysed with multiple one-way ANOVA using a false discovery rate (FDR) value of 0.1 and Fisher's least significant difference test post hoc analyses $Q = 10\%$ ($^*Q < 0.1$; exact *Q* values are in the Source Data). Tumour studies are from two independent cohorts.

Extended Data Fig. 5b). Reconstitution of *NDUFS2* cDNA restored the OCR, pyruvate prototrophy and in vivo tumour growth (Extended Data Fig. 5c–g). Mitochondrial complex I has two key functions: (1) donating electrons from NADH to ubiquinone to result in the generation of NAD+, which allows the oxidative TCA cycle to function, and (2) proton-pumping, which contributes to the generation of ATP

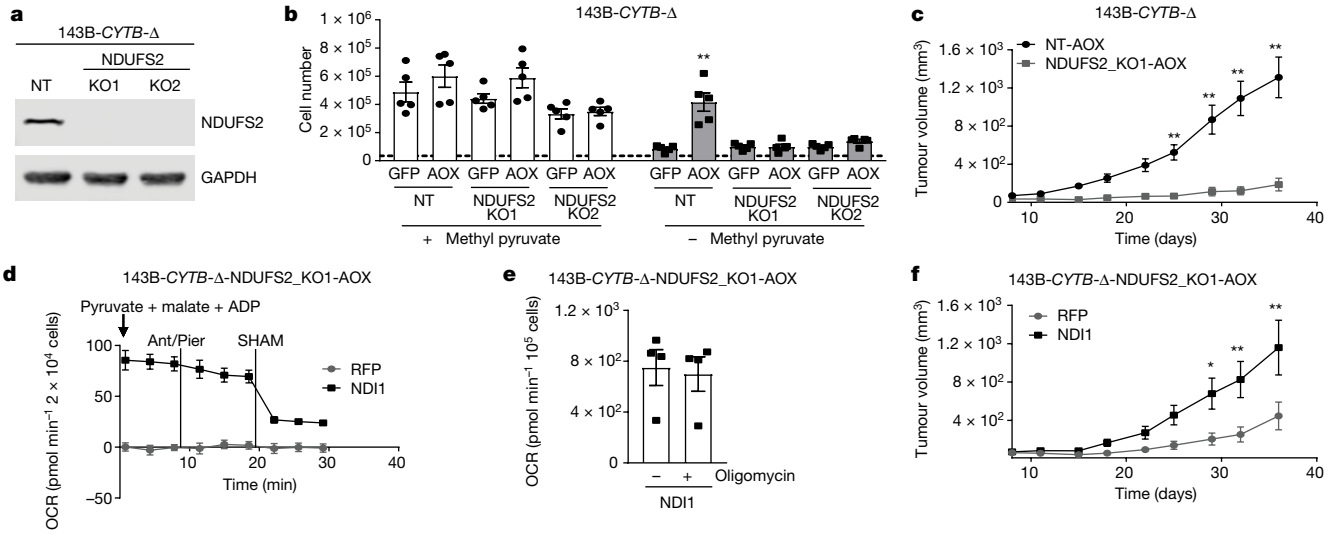


Fig. 3 | Complex I is necessary for tumour growth. **a**, Western blot analysis of NDUFS2 protein levels in 143B-*CYTB*-Δ non-targeting (NT) and 143B-*CYTB*-Δ-NDUFS2_KO cells. GAPDH was used as a loading control. Representative of two independent experiments. **b**, 143B-*CYTB*-Δ-NT and 143B-*CYTB*-Δ-NDUFS2_KO cells expressing either GFP or AOX were grown in medium containing uridine and in the presence or absence of methyl pyruvate, and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **c**, Average tumour volume of xenografts from 143B-*CYTB*-Δ-NT-AOX and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX cells ($n = 10$ mice). **d**, Complex I-driven OCR of permeabilized 143B-*CYTB*-Δ-NDUFS2_KO1 cells expressing AOX and either RFP or NDI1. Piericidin A (Pier; 1 μM) and antimycin A (Ant; 1 μM) were used to inhibit complex I and III, respectively. Salicylhydroxamic acid (SHAM; 2 mM) was used to inhibit AOX activity ($n = 6$ biologically independent experiments). **e**, OCR in the presence or absence of oligomycin in 143B-*CYTB*-Δ-NDUFS2_KO1 cells expressing AOX and NDI1 ($n = 4$ biologically independent experiments). **f**, Average tumour volume of xenografts from 143B-*CYTB*-Δ-NDUFS2_KO1 cells expressing AOX and either RFP or NDI1 ($n = 10$ mice). Data are mean ± s.e.m. (**b**–**f**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**e**) or two-way ANOVA (**b**, **c**, **f**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). For gel source data, see Supplementary Fig. 2. Tumour studies are from two independent cohorts.

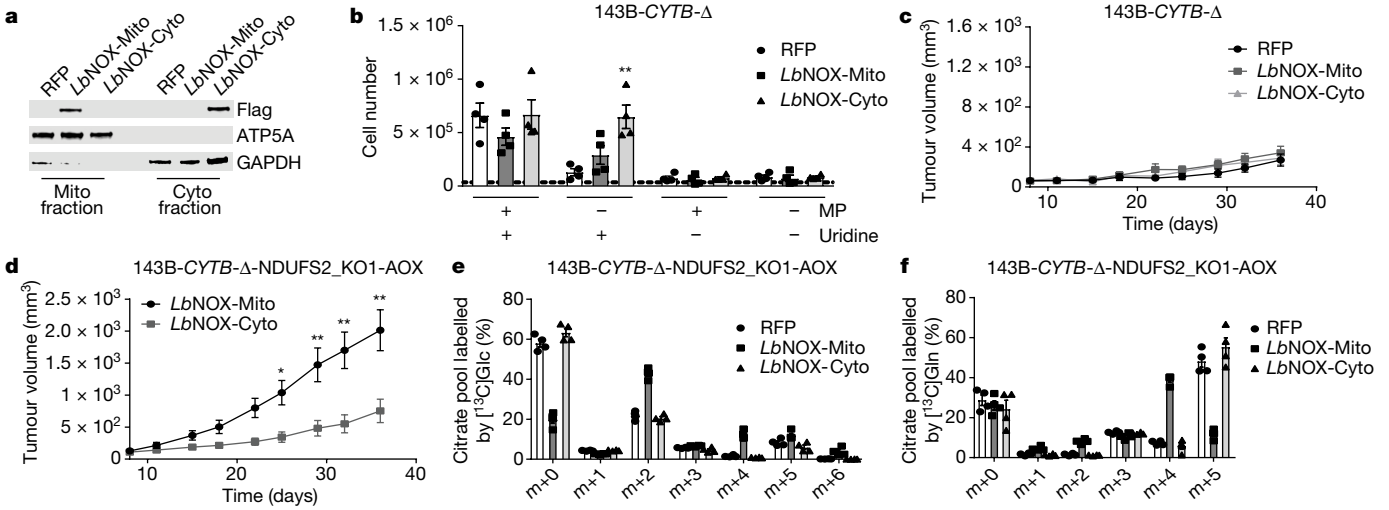


Fig. 4 | Mitochondrial NAD+ regeneration is necessary but not sufficient for tumour growth in vivo. a, Subcellular localization of *Lb*NOX in 143B-*CYTB*-Δ cells determined by cell fractionation. ATP5A is a mitochondrial marker and GAPDH is a cytosolic marker. Representative of three independent experiments. **b**, 143B-*CYTB*-Δ cells expressing mitochondrial (Mito) or cytosolic (Cyto) *Lb*NOX, or red fluorescent protein (RFP) control were grown in the presence or absence of methyl pyruvate and/or uridine, and cell number was assessed after 72 h ($n = 4$ biologically independent experiments). **c**, Average tumour volume of xenografts from 143B-*CYTB*-Δ-RFP, 143B-*CYTB*-Δ-*Lb*NOX-Mito and 143B-*CYTB*-Δ-*Lb*NOX-Cyto cells ($n = 9$ mice). **d**, Average tumour volume of xenografts from 143B-*CYTB*-Δ-NDUFS2_KO1 cells expressing AOX and either mitochondrial or cytosolic *Lb*NOX ($n = 10$ mice). **e**, **f**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX cells expressing either RFP or *Lb*NOX in mitochondria or cytosol were labelled for 6 h with [U-^{13}C]glucose (**e**) or [U-^{13}C]glutamine (**f**) in the presence of methyl pyruvate, and the percentage of labelled citrate pools was examined. $m + 0$ pools represent unlabelled fractions ($n = 4$ biologically independent experiments). Data are mean ± s.e.m. (**b**–**f**). $^{*}P < 0.05$, $^{**}P < 0.01$, two-way ANOVA (**b**–**d**) with a Bonferroni test for multiple comparisons (exact P values are in the Source Data). For gel source data, see Supplementary Fig. 3. Tumour studies are from two independent cohorts.

through OXPHOS. To investigate whether the proton-pumping activity of complex I is necessary for tumour growth, we ectopically expressed control RFP or the *Saccharomyces cerevisiae* alternative NADH dehydrogenase (NDI1) in the AOX-expressing 143B-*CYTB*-Δ-NDUFS2-knockout (NDUFS2_KO) cells (Extended Data Fig. 6a). NDI1 can oxidize NADH to NAD+ by donating electrons to ubiquinone without generating proton-motive force[14]. Therefore, NDI1 restored mitochondrial NADH oxidation, alleviated pyruvate auxotrophy, and changed the metabolome of the AOX-expressing 143B-*CYTB*-Δ-NDUFS2_KO cells (Fig. 3d, Extended Data Fig. 6b, c). The ETC complex V inhibitor oligomycin did not decrease the OCR, indicating that these cells are unable to conduct OXPHOS (Fig. 3e). Moreover, NDI1-and AOX-expressing 143B-*CYTB*-Δ-NDUFS2_KO cells underwent cell death when glucose was replaced by galactose, which forces cells to rely on mitochondrial ATP for survival (Extended Data Fig. 6d). Notably, NDI1 increased tumour growth of AOX-expressing 143B-*CYTB*-Δ-NDUFS2_KO cells (Fig. 3f, Extended Data Fig. 6e), indicating that OXPHOS is not necessary to support tumour growth.

To test whether mitochondrial NAD+ regeneration is necessary and sufficient for tumour growth, we used the water-forming NADH oxidase from *L. brevis* (*Lb*NOX) targeted to the mitochondrial matrix or cytosol[13]. Expression of the *Lb*NOX increases NAD+/NADH ratios in the respective compartments, and importantly, restores proliferative defects caused by ETC inhibition in vitro[13]. To investigate mitochondrial NAD+ sufficiency for tumour growth, we first expressed the mitochondrial or cytosolic *Lb*NOX in 143B-*CYTB*-Δ cells (Fig. 4a, Extended Data Fig. 7a). Both cytosolic and mitochondrial *Lb*NOX alleviated pyruvate auxotrophy by increasing the NAD+/NADH ratio, levels of TCA cycle metabolites, and cell proliferation in vitro (Fig. 4b, Extended Data Fig. 7b–f). However, neither cytosolic nor mitochondrial *Lb*NOX expression was sufficient to rescue tumour growth in vivo (Fig. 4c, Extended Data Fig. 7g). To further test whether regeneration of NAD+ is necessary for tumour growth, we expressed the mitochondrial or cytosolic *Lb*NOX in AOX-expressing 143B-*CYTB*-Δ NDUFS2-knockout cells (Extended Data Fig. 8a). Expression of either *Lb*NOX relieved the pyruvate auxotrophy of the cells, and increased the NAD+/NADH ratio and aspartate levels of these cells in vitro (Extended Data Fig. 8b–d). Both mitochondrial and cytosolic *Lb*NOX changed the metabolome of the cells in vitro (Extended Data Fig. 6c). Owing to the inability of these cells to perform OXPHOS, cell death was observed when glucose in the growth medium was replaced by galactose (Extended Data Fig. 8e). Notably, only mitochondrial *Lb*NOX supported significant tumour growth in vivo (Fig. 4d, Extended Data Fig. 8f). The expression of both *Lb*NOX oxidases inside the in vivo tumours was confirmed (Extended Data Fig. 8g). A potential difference between cells with mitochondrial versus cytosolic *Lb*NOX is the ability of the former to conduct oxidative TCA metabolism while the latter can only perform reductive TCA metabolism (Extended Data Fig. 9a). It is likely that the oxidative TCA cycle flux generates metabolites more efficiently than reductive TCA cycle flux to support macromolecule synthesis for tumour growth. Indeed, mitochondrial *Lb*NOX supported oxidative TCA cycle flux in the presence of pyruvate, as identified by the increased levels of $m + 2$ and $m + 4$ mass isotopomers of citrate from [U-^{13}C]glucose and [U-^{13}C]glutamine, respectively (Fig. 4e, f). By contrast, cytosolic *Lb*NOX supported reductive metabolism in the presence of pyruvate, as the levels of the $m + 5$ mass isotopomer of citrate, and the $m + 3$ mass isotopomers of fumarate, aspartate and malate from [U-^{13}C]glutamine were significantly increased (Fig. 4f, Extended Data Fig. 9b–d). Notably, the same results were observed in the absence of pyruvate when cells were labelled with [U-^{13}C]glutamine (Extended Data Fig. 9e–h). Collectively, our results indicate that although cytosolic NAD+ regeneration can better rescue the metabolic phenotype of complex I deficient cells in vitro (Extended Data Fig. 6c), mitochondrial NAD+ regeneration, probably owing to its unique ability to restore oxidative TCA cycle flux, is more efficient at supporting tumour growth in vivo. These results further support the limitation of in vitro systems in reflecting the metabolic needs of tumours in vivo.

Oxidation of ubiquinol is also required for mitochondrial complex II function. We tested whether complex II is essential for tumour growth by genetically inactivating *SDHA*, which encodes an essential subunit of complex II, in AOX-expressing 143B-*CYTB*-Δ cells (Extended Data Fig. 10a, b). Previous studies have demonstrated that loss of complex II in cancer cells causes pyruvate auxotrophy for aspartate synthesis[21,22]. Indeed, loss of SDHA in AOX-expressing 143B-*CYTB*-Δ cells diminished complex II activity, induced pyruvate auxotrophy, and suppressed tumour growth in vivo (Extended Data Fig. 10c–f). Reconstitution of SDHA cDNA in the cells rescued those phenotypes (Extended Data Fig. 11a–e). It is important to note that there are rare cancers that exhibit mutations in *SDH* subunits as well as the TCA cycle enzyme fumarate hydratase (*FH*) to generate high levels of succinate and fumarate as oncometabolites[1]. However, these cancer cells are able to conduct reductive TCA cycle metabolism to generate the necessary metabolites for proliferation[1,20–22].

Our results indicate that mitochondrial complex III function is required for tumour growth. Complex III is necessary for ubiquinol oxidation which is essential for complex I and II function and for the de novo pyrimidine synthesis pathway. Our findings indicate that complexes I and II are required for tumour growth owing to the regeneration of mitochondrial NAD+ and FAD, which enable oxidative TCA cycle flux. Cancer cells use various mechanisms including glutaminolysis and autophagy to replenish TCA cycle metabolites[23]. As a result, inhibition of glutaminolysis or autophagy in certain cancers diminishes tumour growth[24,25]. Recently, a study used positron emission tomography (PET) imaging of a radiotracer, 4-[^{18}F]fluorobenzyl triphenylphosphonium (^{18}FBnTP), to non-invasively measure the mitochondrial membrane potential of lung tumours in vivo, which was predictive of their response to complex I inhibitors[26]. In the future, it will be of interest to determine whether complex I inhibitors[27] or mitochondrial TCA cycle inhibitors[28] are efficacious in ongoing phase 3 clinical trials, and to develop safe but potent chemotherapeutic strategies that target mitochondrial metabolism in combination with upcoming technologies that profile the metabolic states of human cancers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2475-6.

1. DeBerardinis, R. J. & Chandel, N. S. We need to talk about the Warburg effect. *Nat. Metab.* **2**, 127–129 (2020).
2. Weinberg, F. et al. Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl Acad. Sci. USA* **107**, 8788–8793 (2010).
3. Tan, A. S. et al. Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. *Cell Metab.* **21**, 81–94 (2015).
4. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, (2014).
5. Kuntz, E. M. et al. Targeting mitochondrial oxidative phosphorylation eradicates therapy-resistant chronic myeloid leukemia stem cells. *Nat. Med.* **23**, 1234–1240 (2017).
6. Roth, K. G., Mambetsariev, I., Kulkarni, P. & Salgia, R. The mitochondrion as an emerging therapeutic target in cancer. *Trends Mol. Med.* **26**, 119–134 (2020).
7. Navarro, P. et al. Targeting tumor mitochondrial metabolism overcomes resistance to antiangiogenics. *Cell Rep.* **15**, 2705–2718 (2016).
8. Kim, S. H. et al. Phenformin inhibits myeloid-derived suppressor cells and enhances the anti-tumor activity of PD-1 blockade in melanoma. *J. Invest. Dermatol.* **137**, 1740–1748 (2017).
9. Viale, A. et al. Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. *Nature* **514**, 628–632 (2014).
10. Maher, E. A. et al. Metabolism of [U-^{13}C]glucose in human brain tumors in vivo. *NMR Biomed.* **25**, 1234–1244 (2012).
11. Hensley, C. T. et al. Metabolic heterogeneity in human lung tumors. *Cell* **164**, 681–694 (2016).
12. Hakkaart, G. A., Dassa, E. P., Jacobs, H. T. & Rustin, P. Allotopic expression of a mitochondrial alternative oxidase confers cyanide resistance to human cell respiration. *EMBO Rep.* **7**, 341–345 (2006).
13. Titov, D. V. et al. Complementation of mitochondrial electron transport chain by manipulation of the NAD+/NADH ratio. *Science* **352**, 231–235 (2016).
14. Martínez-Reyes, I. et al. TCA cycle and mitochondrial membrane potential are necessary for diverse biological functions. *Mol. Cell* **61**, 199–209 (2016).
15. Garcia-Bermudez, J. et al. Aspartate is a limiting metabolite for cancer cell proliferation under hypoxia and in tumours. *Nat. Cell Biol.* **20**, 775–781 (2018).
16. Sullivan, L. B. et al. Supporting aspartate biosynthesis is an essential function of respiration in proliferating cells. *Cell* **162**, 552–563 (2015).
17. Sullivan, L. B. et al. Aspartate is an endogenous metabolic limitation for tumour growth. *Nat. Cell Biol.* **20**, 782–788 (2018).
18. Birsoy, K. et al. An essential role of the mitochondrial electron transport chain in cell proliferation is to enable aspartate synthesis. *Cell* **162**, 540–551 (2015).
19. Bajzikova, M. et al. Reactivation of dihydroorotate dehydrogenase-driven pyrimidine biosynthesis restores tumor growth of respiration-deficient cancer cells. *Cell Metab.* **29**, 399–416 (2019).
20. Mullen, A. R. et al. Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature* **481**, 385–388 (2011).
21. Lussey-Lepoutre, C. et al. Loss of succinate dehydrogenase activity results in dependency on pyruvate carboxylation for cellular anabolism. *Nat. Commun.* **6**, 8784 (2015).
22. Cardaci, S. et al. Pyruvate carboxylation enables growth of SDH-deficient cells by supporting aspartate biosynthesis. *Nat. Cell Biol.* **17**, 1317–1326 (2015).
23. DeBerardinis, R. J. & Chandel, N. S. Fundamentals of cancer metabolism. *Sci. Adv.* **2**, e1600200 (2016).
24. Guo, J. Y. et al. Activated Ras requires autophagy to maintain oxidative metabolism and tumorigenesis. *Genes Dev.* **25**, 460–470 (2011).
25. Romero, R. et al. Keap1 loss promotes Kras-driven lung cancer and results in dependence on glutaminolysis. *Nat. Med.* **23**, 1362–1368 (2017).
26. Momcilovic, M. et al. In vivo imaging of mitochondrial membrane potential in non-small-cell lung cancer. *Nature* **575**, 380–384 (2019).
27. Molina, J. R. et al. An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nat. Med.* **24**, 1036–1046 (2018).
28. Alistar, A. et al. Safety and tolerability of the first-in-class agent CPI-613 in combination with modified FOLFIRINOX in patients with metastatic pancreatic cancer: a single-centre, open-label, dose-escalation, phase 1 trial. *Lancet Oncol.* **18**, 770–778 (2017).

Methods

Cell culture and drug treatment

143B-*CYTB*-WT and 143B-*CYTB*-Δ cells were previously described[2]. Mouse *Kras*$^{G12D/+}$*p53*$^{-/-}$ (KP) lung tumour cells expressing luciferase were generously provided by T. Papagiannakopoulos. Cells were grown in DMEM containing 4.5 g l^{-1} glucose, 4 mM L-glutamine (Gibco; 11965-126) supplemented with 10% Nu-serum IV (Corning), 1 mM methyl pyruvate, 400 μM uridine, 1% HEPES and 1% antibiotic-antimycotic (Gibco) at 37 °C with 5% CO_2. Hygromycin (600 μg ml^{-1}) was used to select luciferase expressing KP cells. 143B-*CYTB*-WT were treated with: 500 nM antimycin A (Sigma) and 500 nM piericidin A (Sigma). pWPI-EF1-GFP vectors with AOX and NDI1 were a gift from E. Dufour. The full-length coding sequences of *N*DI1, NDUFS2, SDHA, DHODH, *Lb*NOX-Mito (Addgene; 74448; V. Mootha laboratory), and *Lb*NOX-Cyto (Addgene; 75285; V. Mootha laboratory) were subcloned into the pLV-EF1-RFP vector (VectorBuilder). The resultant NDI1, NDUFS2, SDHA, DHODH, *Lb*NOX-Mito and *Lb*NOX-Cyto vectors, as well as their empty vector control, pLV-EF1-RFP, were transfected into 293T cells (ATTC) along with pMD2.G and psPAX2 packaging vectors using jetPRIME transfection reagent (Polyplus) to produce control-RFP, NDI1-RFP, NDUFS2-RFP, SDHA-RFP, DHODH-RFP, *Lb*NOX-Mito-RFP and *Lb*NOX-Cyto-RFP lentivirus. Similarly, AOX vector and its empty vector control, pWPI-EF1-GFP, were transfected into 293T to generate control-GFP and AOX-GFP lentivirus. Three days after transduction with the indicated virus, GFP- or RFP-positive cells were sorted using a BD FACSAria cell sorter. The cells were periodically sorted to maintain high levels of protein expressions. To generate 143B-*CYTB*-Δ-NDUFS2_KO, 143B-*CYTB*-Δ-SDHA_KO, 143B-*CYTB*-Δ-DHODH_KO and KP-QPC_KO cell lines, gene-specific single-guides RNAs (sgRNAs) listed in Supplementary Table 1 were cloned into the pSpCas9(BB)-2A-GFP (PX458) plasmid (Addgene; 48138; F. Zhang laboratory). These sgRNA-Cas9-GFP vectors were transfected into 143B-*CYTB*-Δ or KP cells using the jetPRIME (Polyplus). Two days after transfection, the GFP-positive cells were single-cell sorted into a 96-well plate using a BD FACSAria cell sorter. The cells were grown for 2–3 weeks, and the resultant clonal cell lines were expanded. Immunoblotting was used to confirm knockout of the targeted gene. Cells have not been authenticated. Cells tested negative for mycoplasma contamination.

Cellular fractionation and immunoblot analysis

Mitochondria were isolated with the human mitochondria isolation kit (Miltenyi Biotec) using 7×10^6–9×10^6 cells, following the manufacturer's instructions. Purified mitochondria were then lysed using the 1× cell lysis buffer (Cell Signaling) containing the Halt protease inhibitor cocktail (Thermo Scientific). The cytosolic fraction was prepared by differential centrifugation using cells from one 10 cm dish. Cells were resuspended in 0.5 ml of PBS with the Halt protease inhibitor cocktail, and lysed by passing through a 27.5-guage needle 12 times. Intact cells and nuclei were removed by centrifugation for 10 min at 800*g* at 4 °C. The supernatant was transferred to a new Eppendorf tube and centrifuged for 10 min at 8,000*g* at 4 °C. The resultant supernatant was the cytosolic fraction. Whole-cell lysate extracts were prepared from the indicated cell lines by collecting and lysing cells in 1× cell lysis buffer containing the Halt protease inhibitor cocktail. The Pierce BCA Protein Assay kit (Thermo Scientific) was used to quantify the protein concentrations. Approximately 50–100 μg of lysate was resolved on a SDS–PAGE gel (Bio-Rad) and transferred to a nitrocellulose membrane using the Trans-Blot Turbo Transfer System (Bio-Rad). Membranes were first blocked in 5% milk for 1 h, then incubated in the primary antibody overnight. Primary antibodies used were: anti-NDUFS2 (Abcam; ab103024; 1:500 dilution), anti-QPC (Abcam; ab136679; 1:500 dilution), anti-SDHA (MitoScience; MS204; clone 2E3GC12FB2AE2; 1:500 dilution), anti-DHODH (Santa Cruz; sc-166348; clone E-8; 1:500 dilution), anti-FLAG (Sigma; F1804; clone M2; 1:1,000 dilution), anti-GAPDH (Santa Cruz; sc-32233; clone 6C5, and Sigma; G9545; 1:2,000 dilution), anti-ATP5A (Mitosciences; MS507; clone 15H4C4; 1:1,000 dilution), anti-tubulin (Cell Signaling; 2144; 1:1,000 dilution) and anti-β-actin (Sigma; A2228-100UL; clone AC-74,). IRDye 800CW goat anti-rabbit (LI-COR; 926-32211) and IRDye 680RD goat anti-mouse (LI-COR; 926-68070) were used as secondary antibodies. Image Studio Lite version 3.1 (LI-COR) was used for the analysis of protein levels.

Mouse models and tumour studies

Uqcrq (QPC) floxed (flox), wild-type (WT) and null (−) alleles were genotyped using the following primers: QPC-F- CTTCCGCTCCTCCCGGAAGT; QPC-R- TTCCCAAACTCGCGGCCCATG and QPC-null- CAATTCCAGCCAACAGTCCC. *Ubc-cre*^{ERT2} mice were obtained from the Jackson Laboratory. *Uqcrq*$^{flox/flox}$, *Uqcrq*$^{WT/-}$ and *Ubc-cre*^{ERT2} mice were crossed to generate T-ALL donors containing *Ubc-cre*^{ERT2} alleles with floxed/null *Uqcrq* (*Uqcrq*$^{flox/-}$*;Ubc-cre*^{ERT2}), or wild-type/null *Uqcrq* (*Uqcrq*$^{WT/-}$*; Ubc-cre*^{ERT2}) as control. Mice of both sexes aged 8–12 weeks old were used for experiments. Mice were not randomized to experimental groups, but were age-matched, sex-matched, and littermates when possible. For xenograft tumour studies, 4×10^6 cells were subcutaneously injected into male J:Nu mice (8–12 weeks). Tumours were measured twice a week and tumour volume was calculated using the following equation: $(4/3) \times \pi \times$ (arithmetic mean of 2 calliper measurements)/2)^3. At the completion of the study, mice were euthanized and the tumours were extracted and weighed. Mice were euthanized before the endpoint was reached if tumours reached 2 cm diameter, developed ulcerations or mice exhibited distress. For the orthotopic lung tumour model, 2.5×10^5 KP cells in 50–75 μl of PBS plus 2.5mM EDTA were intratracheally instilled in C57BL/6J mice as previously described[29]. In vivo luciferase was imaged on IVIS or LAGO system to monitor tumour growth. The fur on the chest was first removed using Nair hair removal cream. Subsequently, 150 μl of RediJect D-Luciferin Ultra Bioluminescent Substrate (PerkinElmer) was injected inraperitoneally, and images were taken after 10 min. Images were processed using the Living Image or Aura software to measure the background-corrected bioluminescence signal from the tumours. Mice were euthanized by 20 weeks after tumour administration, or after losing 15–20% initial weight or displaying overt distress. All mice were housed in the Northwestern University animal vivarium and we have complied with all relevant ethical regulations in accordance with Northwestern University Institutional Animal Care and Use Committee (IACUC).

Bone marrow isolation and leukaemic transformation

Platinum-E retroviral packaging cells and MIGR1-Notch1$^{\Delta E}$-GFP vector were a gift from P. Ntziachristos. Platinum-E cells were transfected with MIGR1-Notch1$^{\Delta E}$-GFP plasmid using jetPRIME (Polyplus) in order to generate the Notch1$^{\Delta E}$-GFP retrovirus. Bone marrow cells were obtained from *Uqcrq*$^{flox/-}$*;Ubc-cre*^{ERT2} and *Uqcrq*$^{WT/-}$*;Ubc-cre*^{ERT2} donor mice by grinding the pelvis, femur and tibia bones with a mortar and a pestle. From the bone marrow cells, HSCs were isolated by a CD117$^+$ positive selection magnetic bead isolation kit (StemCell). Subsequently, HSCs were transduced with the Notch1$^{\Delta E}$-GFP retrovirus by centrifugation at 25 °C at 1,500*g* for 90 min, followed by incubation in 37 °C overnight. The virus was removed the next day morning, after which cells were allowed to rest for 2 days. The transduction and culturing of HSCs were performed in Opti-MEM (Thermo Fisher) supplemented with 10 ng ml^{-1} IL-3 (PeproTech), 10 ng ml^{-1} IL-7 (PeproTech), 50 ng ml^{-1} SCF (PeproTech), 50 ng ml^{-1} FLT3L (PeproTech), and 20 ng ml^{-1} IL-6 (PeproTech). Approximately 24 h before adoptive transfer, wild-type C57BL/6J recipients were lethally irradiated at approximately 1,000 rad. On the day of the transfer, lineage (CD4, CD8a, B220, CD11b, Gr-1, NK1.1, Ter-119)-negative and GFP-positive cells were sorted on BD FACS Aria systems. Antibodies used were: anti-Mouse Ter-119 (eBioscience; 48-5921-80; clone TER-119), anti-Mouse NK1.1 (eBioscience; 48-5941-80; clone PK136,), anti-Human/Mouse CD45R (B220) (eBioscience;

48-0452-80; clone RA3-6B2), anti-Mouse CD8a (eBioscience; #48-0081-82; clone 53-6.7), anti-Mouse CD11b (eBioscience; 48-0112-80; clone M1/70,), anti-Mouse Ly-6G (Gr-1) (eBioscience; 48-5931-80; clone RB6-8C5), anti-Mouse CD4 (Tonbo Biosciences; 75-0041-U100; clone GK1.5). Approximately 50,000–100,000 GFP-positive HSCs, along with 500,000 support bone marrow cells isolated from wild-type C57BL/6J mice, were injected intravenously into the recipient mice. At 3 and 4 weeks after transfer, peripheral blood from the recipients was analysed to assess the presence of circulating GFP-positive T-ALL cells. Once the percentage of GFP-positive cells in the peripheral blood reached approximately 5–10%, the recipients were oral gavaged with 320 mg kg^{-1} tamoxifen suspended in corn oil four times, once every 2 days. For up to 25 weeks after tamoxifen administration, the recipients were closely monitored for any signs of malignancy, including weight loss, hunched posture and lethargy. The recipients were euthanized upon displaying 15–20% weight loss, or at up to 30 weeks after transfer. Cell from spleen and bone marrow (from one set of pelvis, femur and tibia) were obtained from each recipient, and stained with Ghost Dye Red 780 (Tonbo Bioscience). The cells were resuspended in FACS buffer (DPBS with 10% NuSerum IV) with PKH reference microbeads (Sigma). The number and the percentage of GFP-positive T-ALL cells were analysed on BD FACSymphony and FlowJo software version 10.4.2.

Proliferation and cell viability analysis

Approximately 3.5×10^4 cells were plated on 6-well plates. Cells were expanded in the presence or absence of methyl pyruvate and/or uridine for 72 h. To assess proliferation, cells were counted using AccuCount Fluorescent Particles (Spherotech) by flow cytometry. Cell viability was determined by measuring the percentage DAPI-positive population by flow cytometry. All flow cytometry assays were performed on BD FACSymphony or BD Fortessa analysers, and data were analysed with the FlowJo software 10.4.2.

Mitochondrial activity studies

The OCR was measured in a XF96 extracellular flux analyser (Seahorse Bioscience). Basal mitochondrial respiration was assessed by subtracting the non-mitochondrial OCR, measured with 1 µM antimycin A and 1 µM piericidin A, from baseline OCR. Coupled respiration was determined by subtracting the OCR in the presence of 1 µM oligomycin A (Sigma) from the basal mitochondrial respiration. To determine mitochondrial complex I activity, growth medium was replaced with mitochondrial assay buffer (70 mM sucrose, 220 mM mannitol, 10 mM KH_2PO_4, 5 mM $MgCl_2$, 2 mM HEPES, 1 mM EGTA, 0.2% (w/v) fatty acid-free BSA, pH 7.2) supplemented with 1 nM Seahorse XF plasma membrane permeabilizer and 10 mM ADP, as well as 2.5 mM malate and 10 mM pyruvate (complex I substrates). To assess mitochondrial complex II activity, growth medium was replaced with the mitochondrial assay buffer containing membrane permeabilizer and ADP, along with 10 mM succinate (complex II substrate) and 1 µM piericidin A, which inhibits the complex I contribution to OCR. The increase in OCR was measured immediately after addition of substrates. Where indicated, Piericidin A and Antimycin A were injected to inhibit complex I and III, respectively. Salicylhydroxamic acid (SHAM; Sigma), was injected to inhibit AOX.

Metabolomics

Subconfluent culture dishes were incubated for 2, 8 or 24 h in DMEM (Gibco; A1443001) supplemented with 15 mM glucose, 2 mM glutamine and 10% dialysed FBS (PEAK Serum), in the presence or absence of 1 mM methyl pyruvate and/or 400 µM uridine. Following the incubation, cells were washed with ice-cold 0.9% NaCl, and overlaid with ultra-cold HPLC grade-methanol/water (80/20, v/v). The plates were incubated at −80 °C for 20 min, after which cells were scraped and collected. The cell suspensions were then centrifuged at 16,000*g* for 15 min at 4 °C. The supernatant was transferred to a new tube and evaporated to dryness using a SpeedVac concentrator (Thermo Savant). Metabolites were reconstituted in 50% acetonitrile in analytical-grade water, vortex-mixed, and centrifuged to remove debris. Samples were analysed by high-performance liquid chromatography and high-resolution mass spectrometry and tandem mass spectrometry (HPLC–MS/MS). Specifically, system consisted of a Thermo Q-Exactive in line with an electrospray source and an Ultimate3000 (Thermo) series HPLC consisting of a binary pump, degasser, and auto-sampler outfitted with a Xbridge Amide column (Waters; dimensions of 4.6 mm × 100 mm and a 3.5 µm particle size). Mobile phase A contained 95% (v/v) water, 5% (v/v) acetonitrile, 10 mM ammonium hydroxide, 10 mM ammonium acetate, pH 9.0; and mobile phase B was 100% acetonitrile. The gradient was as follows: 0 min, 15% A; 2.5 min, 30% A; 7 min, 43% A; 16 min, 62% A; 16.1–18 min, 75% A; 18–25 min, 15% A with a flow rate of 400 µl min^{-1}. The capillary of the ESI source was set to 275 °C, with sheath gas at 45 arbitrary units, auxiliary gas at 5 arbitrary units and the spray voltage at 4.0 kV. In positive/negative polarity switching mode, an m/z scan range from 70 to 850 was chosen and MS1 data were collected at a resolution of 70,000. The automatic gain control (AGC) target was set at 1×10^6 and the maximum injection time was 200 ms. The top five precursor ions were subsequently fragmented, in a data-dependent manner, using the higher energy collisional dissociation (HCD) cell set to 30% normalized collision energy in MS2 at a resolution power of 17,500. Sample volumes of 10 µl were injected. Data acquisition and analysis were carried out by Xcalibur 4.1 software and Tracefinder 4.1 software, respectively (both from Thermo Fisher Scientific). The peak area for each detected metabolite was normalized by the total ion current, which was determined by integration of all of the recorded or annotated peaks within the acquisition window. For carbon labelling, isotopic labelling was performed in DMEM (Gibco; A1443001) supplemented with 10% dialysed FBS, and 2 mM L-[U-^{13}C]glutamine or 10 mM D-[U-^{13}C]glucose, in the presence or absence of 1 mM methyl pyruvate. After 6 h of labelling, metabolites were extracted with ultra-cold HPLC grade-methanol/water (80/20, v/v) and analysed as previously described[20]. Metabolite analyses were carried out in MetaboAnalyst 4.0. Peak intensities normalized to total ion current in tables were loaded. Missing and 0 values were replaced with half the minimum positive values in the original data assuming to be the detection limit. For heat maps with two groups, *t*-tests with an FDR cut-off value of 0.1 were used to identify significantly changed metabolites. For heat maps with more than two groups, one-way ANOVA with Fisher's least significant difference post hoc analyses and an FDR cut-off value of 0.1 was used to generate a list of significantly changed metabolites among groups. This list was then plotted as a heat map with euclidean distance measures and ward.D clustering algorithm of metabolites/rows. Within row, *z*-scores for each metabolite were plotted.

Measurement of dihydroorotate and orotate ratio

For the measurement of dihydroorotate and orotate, 5×10^6 cells were seeded in a 100 mm cell culture dish, and incubated with DMEM (Gibco; A1443001) supplemented with 15 mM glucose, 2 mM glutamine, 10% dialysed FBS, 1% HEPES and 1% antibiotic-antimycotic. After 24 h, metabolites were extracted. Cells were washed twice with DPBS, and lysed with 600 µl of HPLC grade-methanol/chloroform (67/33, v/v). The cell lysates were collected, vortexed for 30 s, and incubated in liquid nitrogen for 60 s. Samples were then thawed at room temperature, after which 400 µl of a HPLC grade-chloroform/water (50/50, v/v) was added. The lysates with metabolites were centrifuged at 15,000*g* for 30 min at 4 °C. The supernatant was transferred to a new tube and evaporated to dryness using a SpeedVac concentrator (Thermo Savant). Samples were analysed by HPLC–MS/MS as described above using targeted selected ion monitoring (tSIM) mode.

Statistical analysis

P values were calculated as described in each figure legend using Graphpad Prism 7 (Graphpad Software) and MetaboAnalyst 4.0[30]. Data are

presented as mean ± s.e.m. unless stated otherwise. Numbers of biological replicates are indicated in the figure legends. The investigators were not blinded during experiments and outcome assessments. No statistical method was used to predetermine sample size, and experiments were not randomized.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data from the manuscript are available from the corresponding author on request. Source data are provided with this paper.

29. DuPage, M., Dooley, A. L. & Jacks, T. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat. Protoc.* **4**, 1064–1072 (2009).
30. Chong, J. et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46** (W1), W486–W494 (2018).

Acknowledgements This work was supported by the NIH (5R35CA197532) to N.S.C.; a postdoctoral fellowship by Ramon Areces Foundation of Spain to I.M.R.; NCI (T32 CA009560) and NIH (2T32HL076139-16) to H. Kong; NIH (5 T32 CA 9560-33) and Northwestern University Pulmonary department's Cugell fellowship to K.V.; NIH (T32CA09560 and T32HL076139-13) to G.S.M., and NIH (T32 T32HL076139) to S.E.W. E.M.S. is a Cancer Research Institute Irvington Fellow supported by the Cancer Research Institute. We thank Robert H. Lurie Cancer Center Flow Cytometry facility supported by NCI CCSG P30 CA060553 for their invaluable assistance. Imaging work was performed at the Northwestern University Center for Advanced Microscopy generously supported by NCI CCSG P30 CA060553 awarded to the Robert H Lurie Comprehensive Cancer Center. We thank T. Papagiannakopoulos and S. Leboeuf for providing KP cells. We thank C. Moraes for *CYTB*-Δ cells. Original *Lb*NOX constructs were generated by V. Mootha's laboratory and acquired from Addgene. We thank E. Dufour and H. Jacobs for original NDI1 and AOX constructs. We thank P. Ntziachristos for the MIGR1-Notch1$^{\Delta E}$-GFP vector. We thank the members of the Chandel laboratory for discussions.

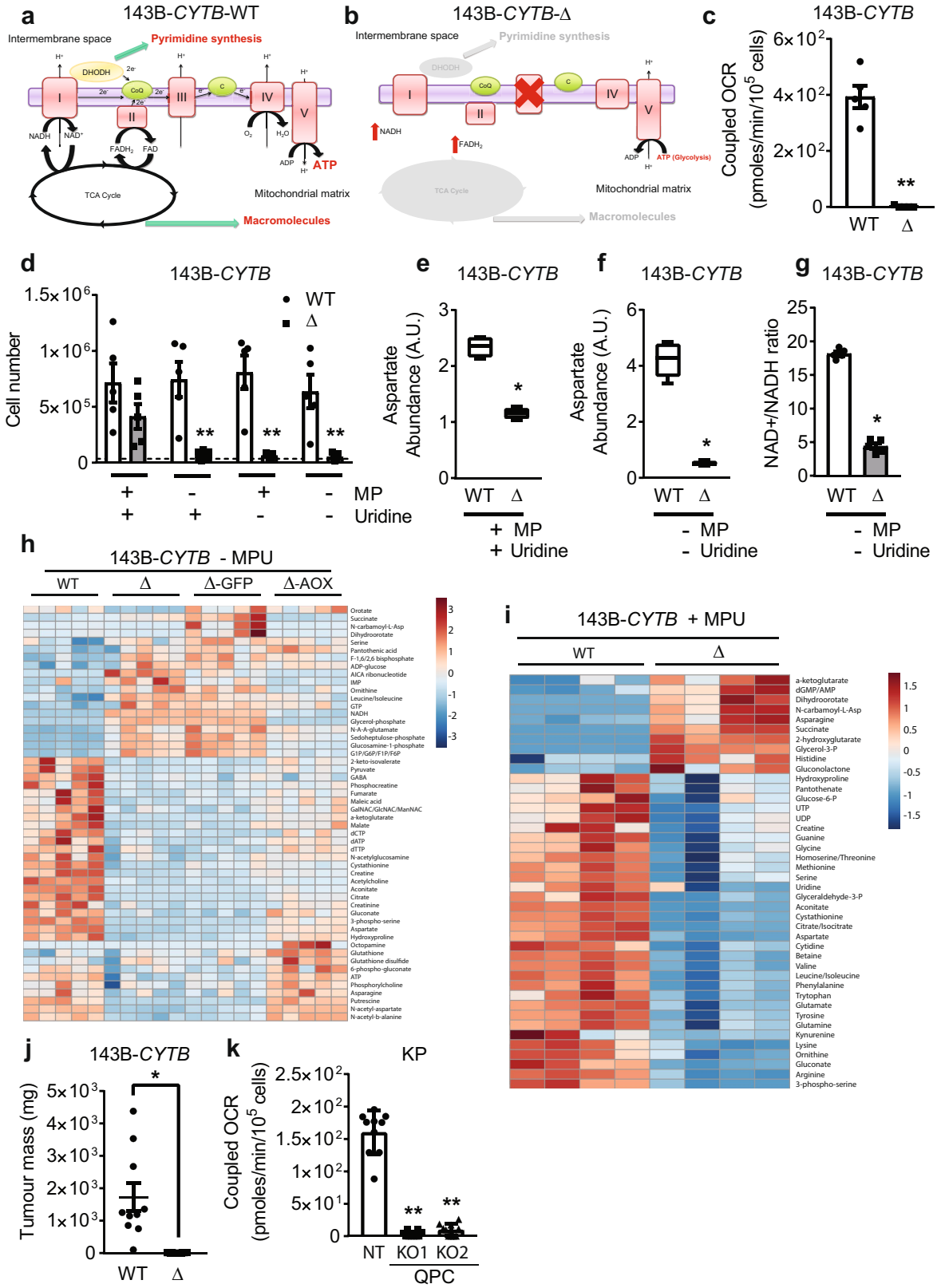
Author contributions I.M.-R. and N.S.C. conceptualized the study, interpreted the data, and wrote the manuscript with the input of co-authors. I.M.-R., L.R.C., H. Kong, M.W., H. Kihshen, G.S.M., C.R.R. and K.V. carried out most of the experiments in the paper. I.M.-R., L.R.C. and H. Kong performed leukaemia experiments. I.M.-R., L.R.C., H. Kong and K.V. performed lung tumour experiments. I.M.-R., L.R.C., C.R.R, G.S.M. and S.E.W. carried out xenograft experiments. I.M.-R., M.W. and H. Kihshen performed experiments to analyse the OCR in intact and permeabilized cells. I.M.-R., C.R.R. and M.W performed experiments to assess proliferation. I.M.-R. and M.W. performed western blot experiments. I.M.-R., L.R.C., G.S.M. and P.G. conducted and analysed metabolomics and carbon flux experiments. R.P., G.R.S.B., E.M.S. and S.E.W. provided technical expertise with mouse experiments.

Competing interests N.S.C. is a scientific advisory board member of Rafael Pharmaceuticals.

Additional information

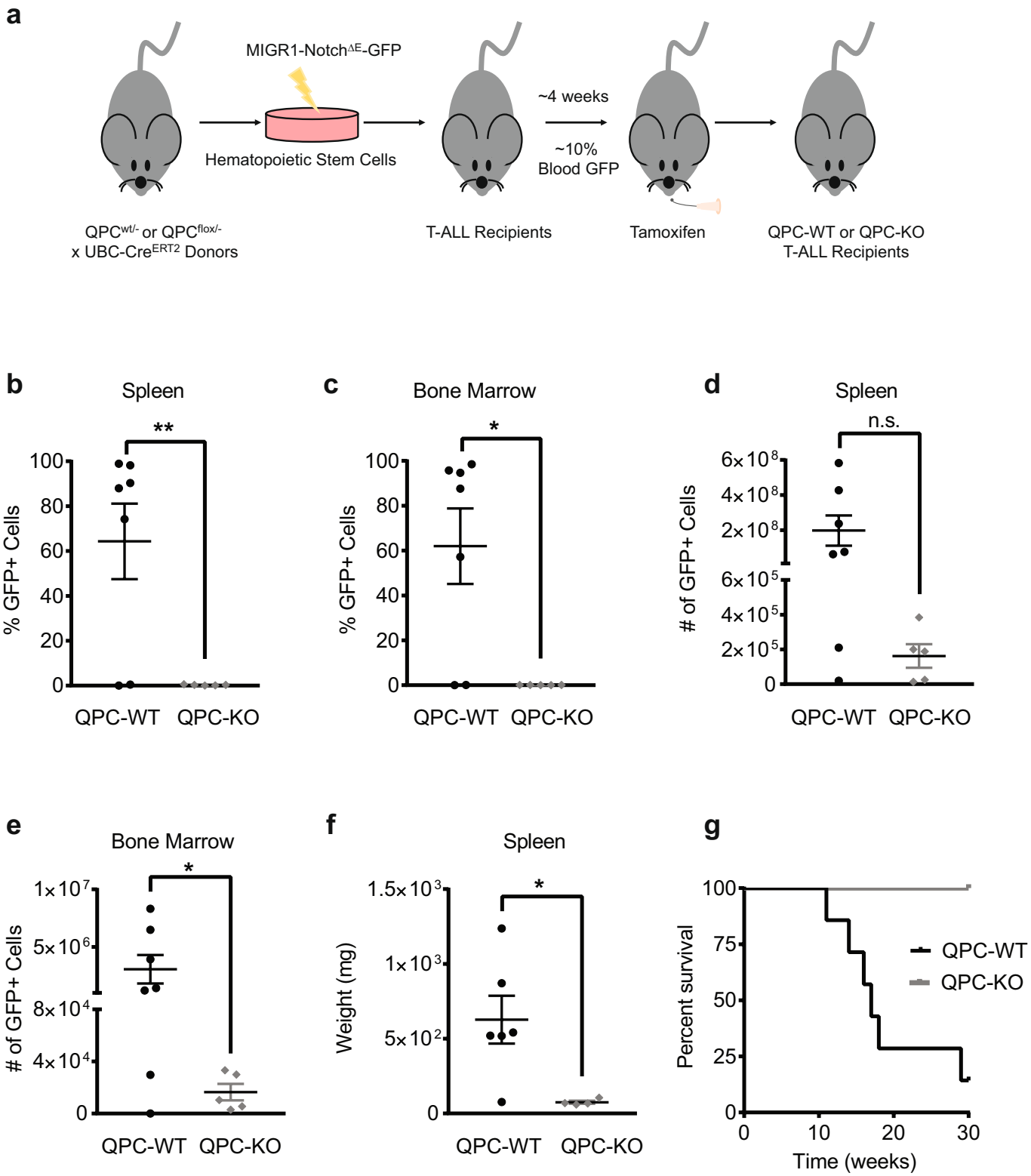
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2475-6.

Correspondence and requests for materials should be addressed to N.S.C.

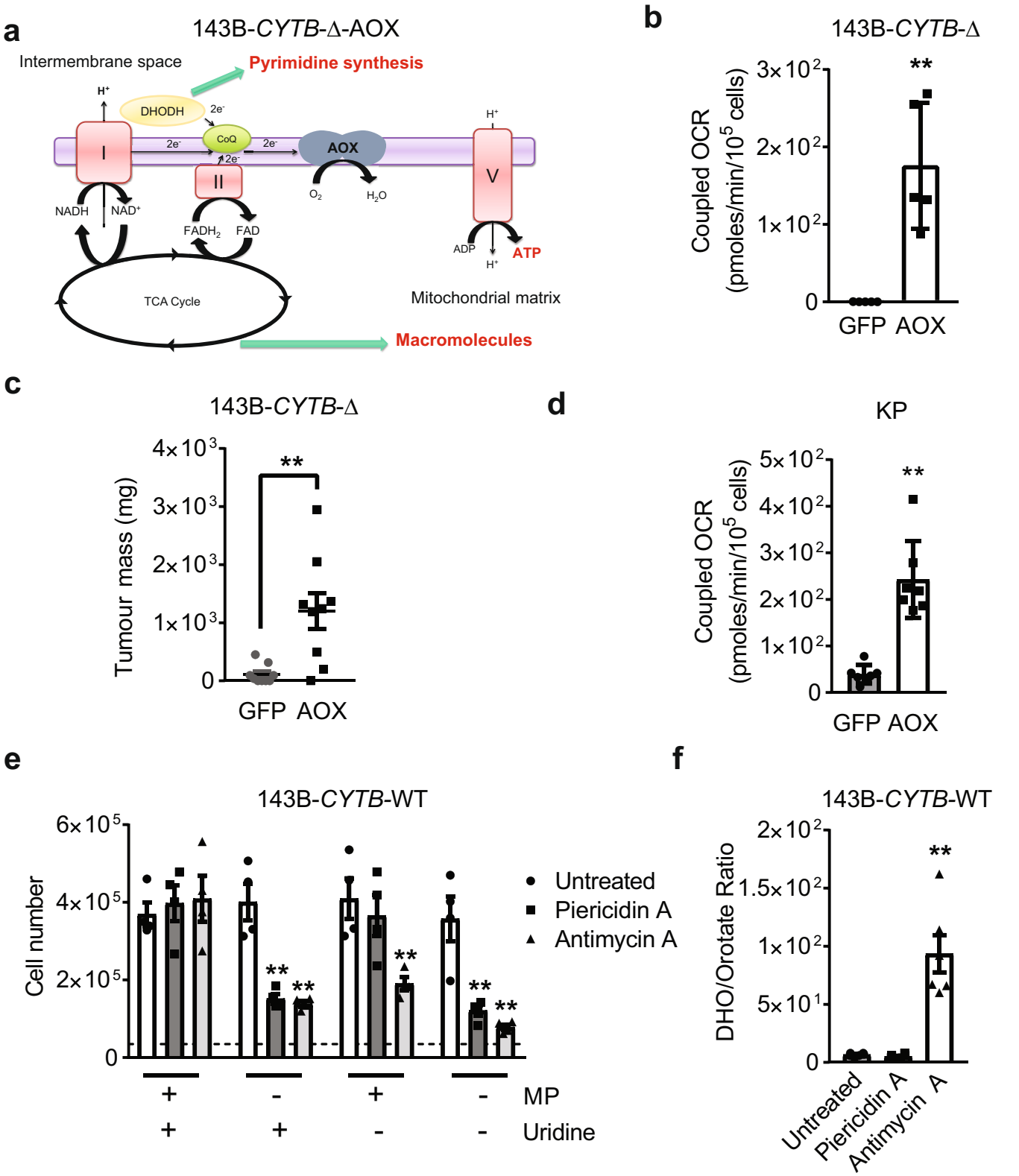


Extended Data Fig. 1 | See next page for caption.

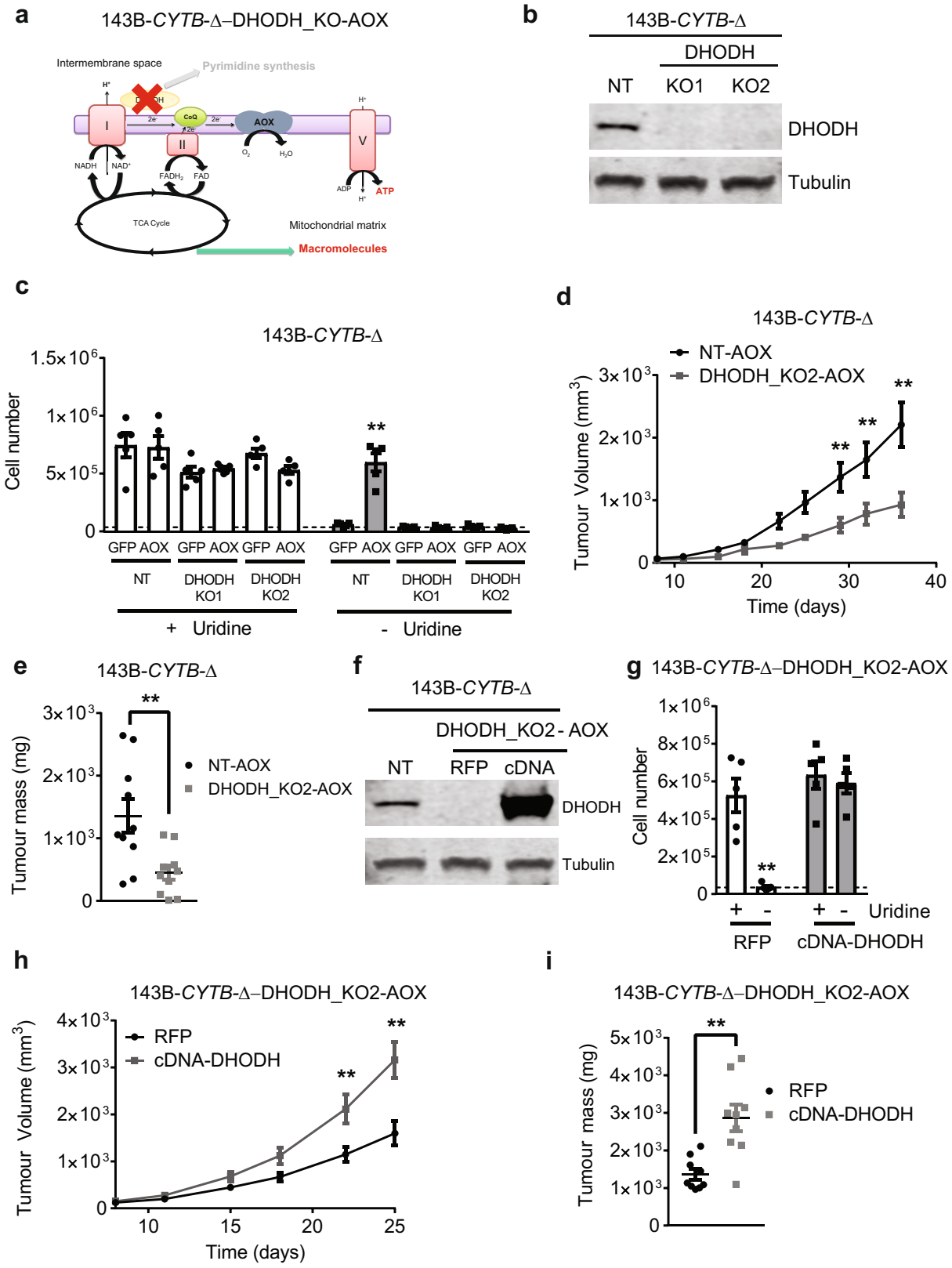
Extended Data Fig. 1 | Metabolite changes in complex III deficient cells in the presence or absence of pyruvate and uridine. a, **b**, Schematic representation of the ETC in 143B-*CYTB*-WT (**a**) and 143B-*CYTB*-Δ cells (**b**). **c**, Coupled OCR of 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 5$ biologically independent experiments). **d**, 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells were grown in the presence or absence of methyl pyruvate and/or uridine and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **e**, Intracellular aspartate levels in the presence of methyl pyruvate and uridine in 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 4$ biologically independent experiments). **f**, Intracellular aspartate levels in the absence of methyl pyruvate and uridine in 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 5$ biologically independent experiments). **g**, Intracellular NAD+/NADH ratio in the absence of methyl pyruvate and uridine of 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 5$ biologically independent experiments). **h**, The heat map displays the relative abundance of significantly changed metabolites in 143B-*CYTB*-WT, 143B-*CYTB*-Δ cells and in 143B-*CYTB*-Δ cells expressing either GFP or AOX in the absence of methyl pyruvate and uridine. A red–blue colour scale depicts the abundance of the metabolites (red: high, blue: low) ($n = 5$ biologically independent experiments). **i**, The heat map displays the relative abundance of significantly changed metabolites in 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells in the presence of methyl pyruvate and uridine ($n = 4$ biologically independent experiments). **j**, Tumour mass of xenografts from 143B-*CYTB*-WT and 143B-*CYTB*-Δ cells ($n = 10$ mice per group from two independent cohorts). **k**, Coupled OCR of KP-NT and KP-QPC_KO cells ($n = 10$ technical replicates from two independent experiments). Data are mean ± s.e.m. (**c**–**g**, **j**) or mean ± s.d. (**k**). $^{*}P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**c**, **g**, **j**), two-way ANOVA (**d**) with a Bonferroni test for multiple comparisons or one-way ANOVA (**k**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). Metabolites levels were analysed with multiple one-way ANOVA using an FDR of 0.1 and Fisher's least significant difference test post hoc analyses $Q = 10\%$. For two-group heat maps, *t*-tests with an FDR cut-off value of 0.1 were used to identify significantly changed metabolites. Each row was analysed individually. ($^{*}Q < 0.1$; exact *Q* values are in the Source Data).



Extended Data Fig. 2 | Mitochondrial complex III is required for T-ALL growth in vivo. a, Schematic representation of the T-ALL experiments. **b**, **c**, Percentage of GFP$^+$ T-ALL cells from the spleen (**b**) or bone marrow (**c**) of QPC-WT and QPC-KO recipients (WT: $n = 7$; KO: $n = 5$ mice). **d**, **e**, The absolute number of GFP$^+$ T-ALL cells from the spleen (**d**) or bone marrow (**e**) of QPC-WT and QPC-KO recipients (WT: $n = 7$; KO: $n = 5$ mice). **f**, Weight of spleens from QPC-WT and QPC-KO recipients (WT: $n = 6$; KO: $n = 4$ mice). **g**, Survival of mice injected with QPC-WT or QPC-KO T-ALL cells (WT: $n = 7$; KO: $n = 4$ mice). Data are mean ± s.e.m. from three independent experiments. $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests with a Welch's correction (exact *P* values are in the Source Data). Survival curves were compared using the log-rank test ($P < 0.0001$). An example of the gating strategy is provided in Supplementary Fig. 7.

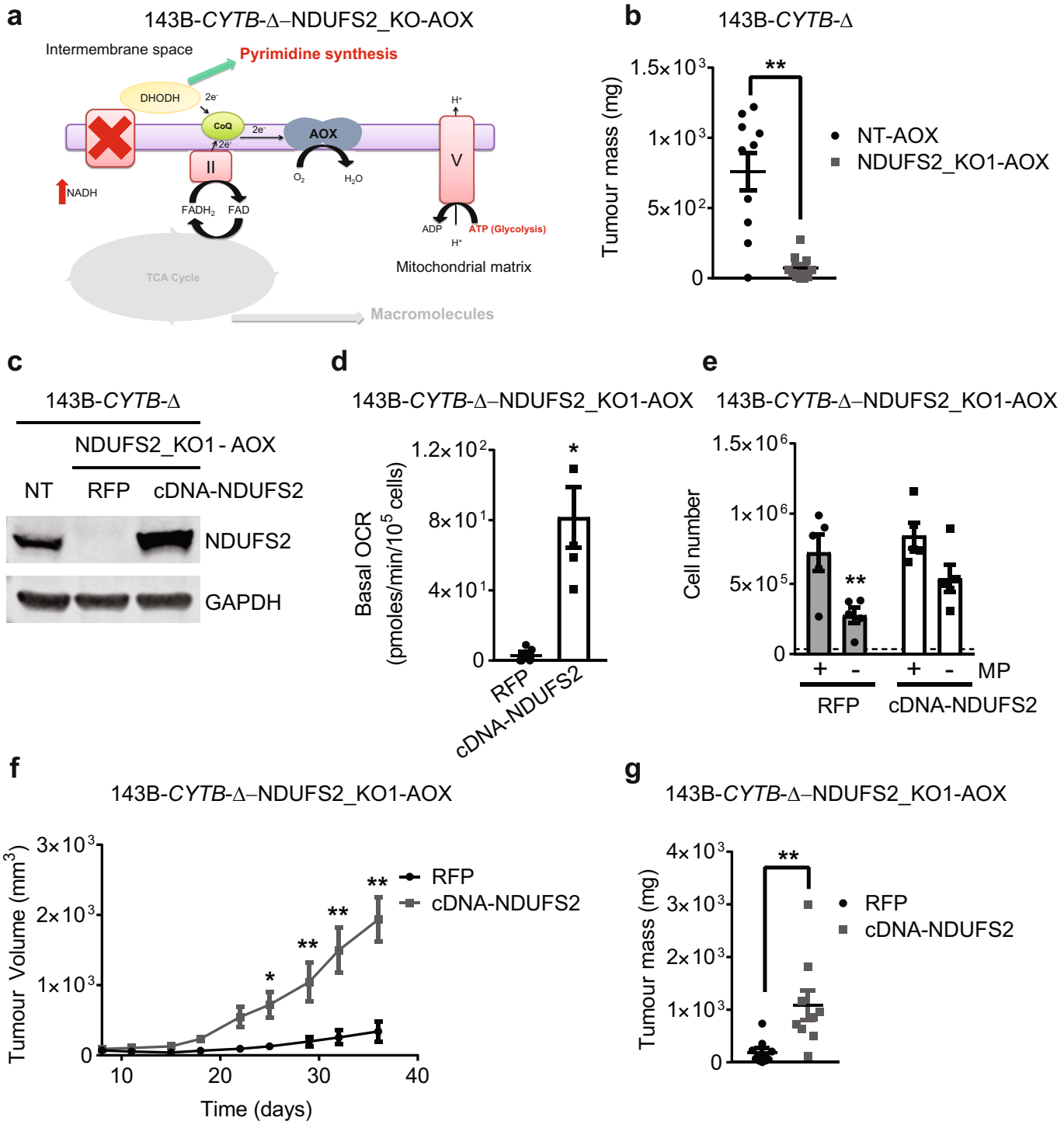


Extended Data Fig. 3 | Complex III-deficient cells are auxotrophic for uridine. a, Schematic representation of the ETC in AOX expressing 143B-*CYTB*-Δ cells. **b**, Coupled OCR of 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells ($n = 5$ biologically independent experiments). **c**, Tumour mass of xenografts from 143B-*CYTB*-Δ-GFP and 143B-*CYTB*-Δ-AOX cells ($n = 9$ mice per group from two independent cohorts). **d**, Coupled OCR of KP-QPC_KO-GFP and KP-QPC_KO-AOX cells ($n = 7$ replicates from one representative of five biologically independent experiments). **e**, 143B-*CYTB*-WT treated or untreated with piericidin A (0.5 µM) or antimycin A (0.5 µM) were grown in the presence or absence of methyl pyruvate and/or uridine and cell number was assessed after 72 h ($n = 4$ biologically independent experiments). **f**, The dihydroorotate-to-orotate ratio was assessed in 143B-*CYTB*-WT treated or untreated with piericidin A (0.5 µM) or antimycin A (0.5 µM) ($n = 6$ biologically independent experiments). Data are mean ± s.e.m. (**b**, **c**, **e**, **f**) or mean ± s.d. (**d**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**b**–**d**), two-way ANOVA (**e**) with a Bonferroni test for multiple comparisons or one-way ANOVA (**f**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data).

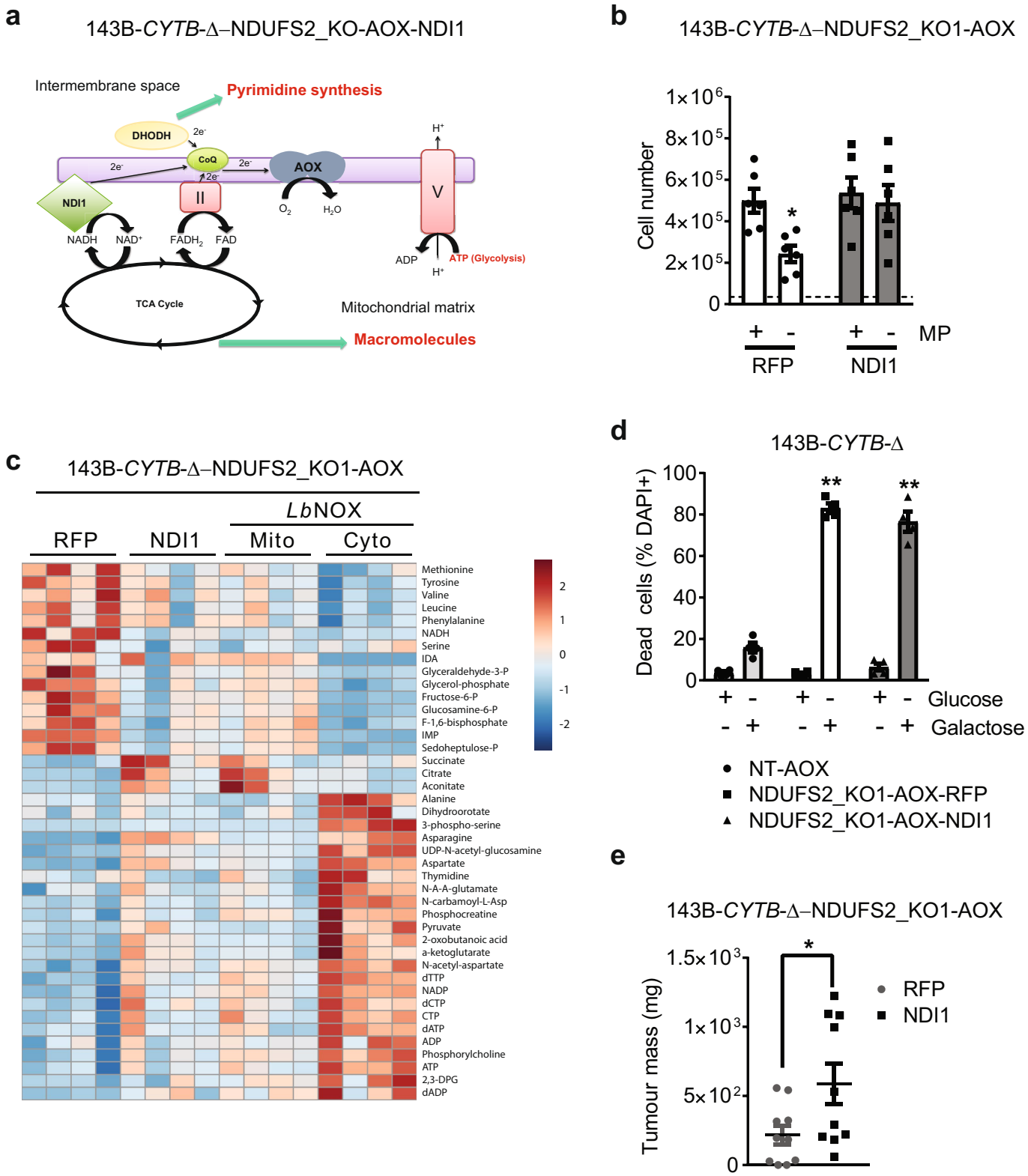


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | De novo pyrimidine synthesis is necessesary for tumour growth. a, Schematic representation of the ETC in 143B-*CYTB*-Δ–DHODH_KO-AOX cells. **b**, Western blot analysis of DHODH in 143B-*CYTB*-Δ non-targeting (NT) and 143B-*CYTB*-Δ-DHODH_KO cells. Tubulin was used as a loading control. Data are representative of two independent experiments. **c**, 143B-*CYTB*-Δ-NT or 143B-*CYTB*-Δ-DHODH-KOs expressing GFP or AOX were grown in the presence or absence of uridine and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **d**, **e**, Average tumour volume (**d**) and tumour mass (**e**) of xenografts from 143B-*CYTB*-Δ-NT-AOX and 143B-*CYTB*-Δ-DHODH_KO2-AOX cells ($n = 10$ mice per group from two independent cohorts). **f**, Western blot analysis of DHODH protein levels in 143B-*CYTB*-Δ-NT, 143B-*CYTB*-Δ-DHODH_KO2-AOX-RFP and 143B-*CYTB*-Δ-DHODH_KO2-AOX-cDNA DHODH cells. Data are representative of three independent experiments. **g**, 143B-*CYTB*-Δ-DHODH_KO2-AOX-RFP and 143B-*CYTB*-Δ-DHODH_KO2-AOX-cDNA DHODH cells were grown in the presence or absence of uridine and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **h**, **i**, Average tumour volume (**h**) and tumour mass (**i**) of xenografts from 143B-*CYTB*-Δ-DHODH_KO2-AOX-RFP and 143B-*CYTB*-Δ-DHODH_KO2-AOX-cDNA DHODH cells ($n = 9$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**c**–**e**, **g**–**i**) $^{*}P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**e**, **i**) or two-way ANOVA (**c**, **d**, **g**, **h**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). For gel source data, see Supplementary Fig. 4.

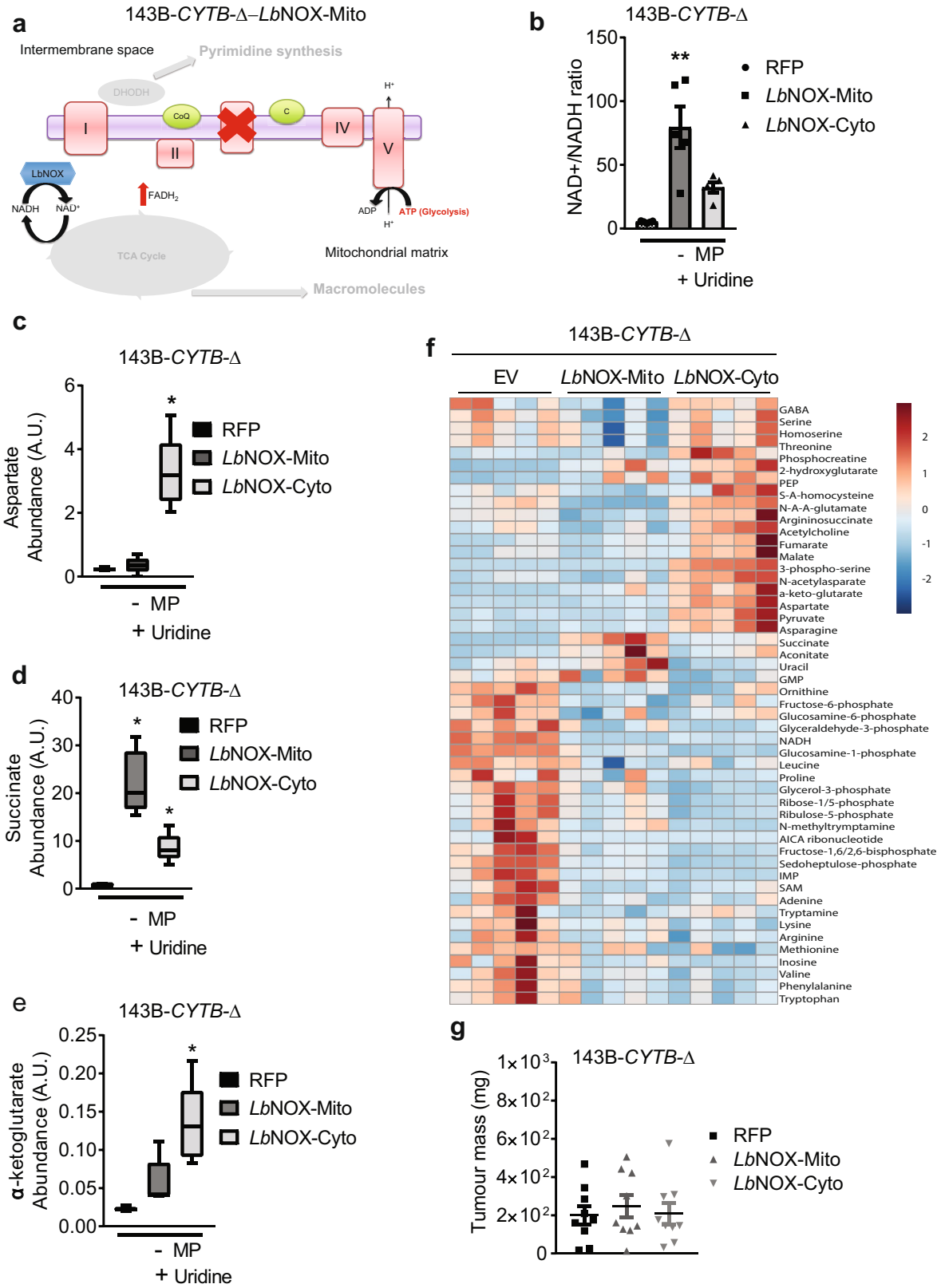


Extended Data Fig. 5 | Restoration of complex I by ectopic expression of *NDUFS2* cDNA rescues tumour growth. a, Schematic representation of the ETC in complex I-deficient 143B-*CYTB*-Δ-NDUFS2_KO-AOX cells. **b**, Tumour mass of xenografts from 143B-*CYTB*-Δ-NT-AOX and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX cells ($n = 10$ mice per group from two independent cohorts). **c**, Western blot analysis of NDUFS2 protein levels in 143B-*CYTB*-Δ-NT cells, and in AOX-expressing 143B-*CYTB*-Δ-NDUFS2_KO1 clone transduced with either RFP or human *NDUFS2* cDNA. GAPDH was used as a loading control. Data representative of two independent experiments. **d**, Basal OCR of AOX expressing 143B-*CYTB*-Δ-NDUFS2_KO1 cells transduced with either RFP or human *NDUFS2* cDNA. **e**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-cDNA NDUFS2 cells were grown in the presence or absence of methyl pyruvate and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **f**, **g**, Average tumour volume (**f**) and tumour mass (**g**) of xenografts from AOX-expressing 143B-*CYTB*-Δ-NDUFS2_KO1 cells transduced with either RFP or human *NDUFS2* cDNA ($n = 9$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**b**, **d**–**g**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**b**, **d**, **g**) or two-way ANOVA (**e**, **f**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). For gel source data, see Supplementary Fig. 2.



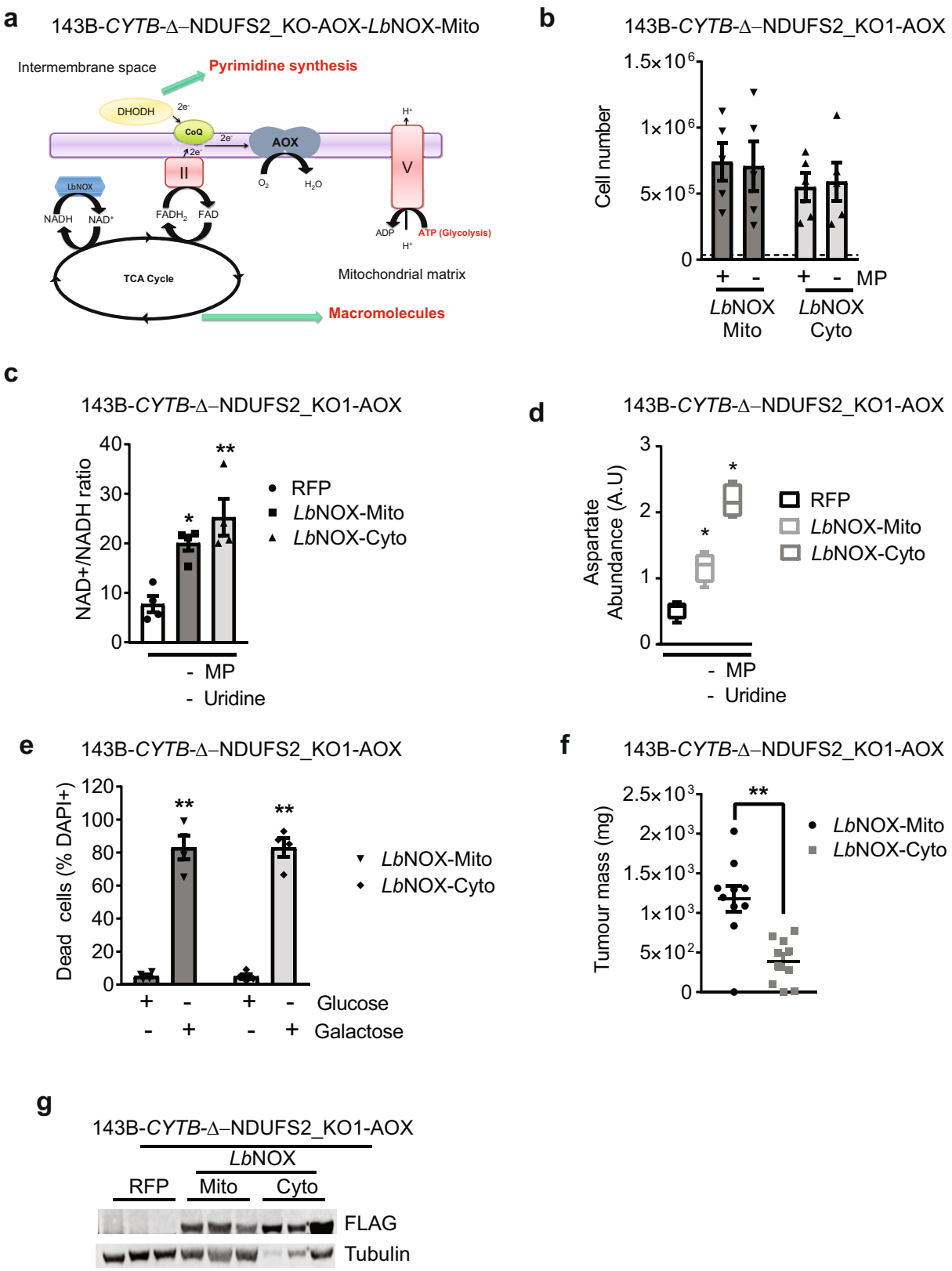
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | NDI1 expression in complex I-deficient cells rescues electron transfer but not ATP production. a, Schematic representation of the ETC in complex I deficient 143B-*CYTB*-Δ-NDUFS2_KO-AOX cells expressing NDI1. **b**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-NDI1 cells were grown in the presence or absence of methyl pyruvate and cell number was assessed after 72 h ($n = 6$ biologically independent experiments). **c**, The heat map displays the relative abundance of significantly changed metabolites in 143B-*CYTB*-Δ–NDUFS2_KO1-AOX cells expressing RFP, NDI1 or *Lb*NOX in either mitochondria or cytosol ($n = 4$ biologically independent experiments). A red–blue colour scale depicts the abundance of the metabolites (red: high, blue: low). Metabolites levels were analysed with multiple one-way ANOVA using an FDR of 0.1 and Fisher's least significant difference test post hoc analyses $Q = 10\%$. Each row was analysed individually. (*$Q < 0.1$; exact Q values are in the Source Data.) **d**, 143B-*CYTB*-Δ-NT-AOX, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-NDI1 cells were grown in media containing 10 mM glucose or 10 mM galactose for 48 h and assessed for cell death ($n = 4$ biologically independent experiments). **e**, Tumour mass of xenografts from 143B-*CYTB*-Δ-NDUFS2_KO1 cells expressing AOX and either RFP or NDI1 ($n = 10$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**b**, **d**, **e**). *$P < 0.05$, **$P < 0.01$, two-tailed t-tests (**e**) or two-way ANOVA (**b**, **d**) with a Bonferroni test for multiple comparisons (exact P values are in the Source Data).



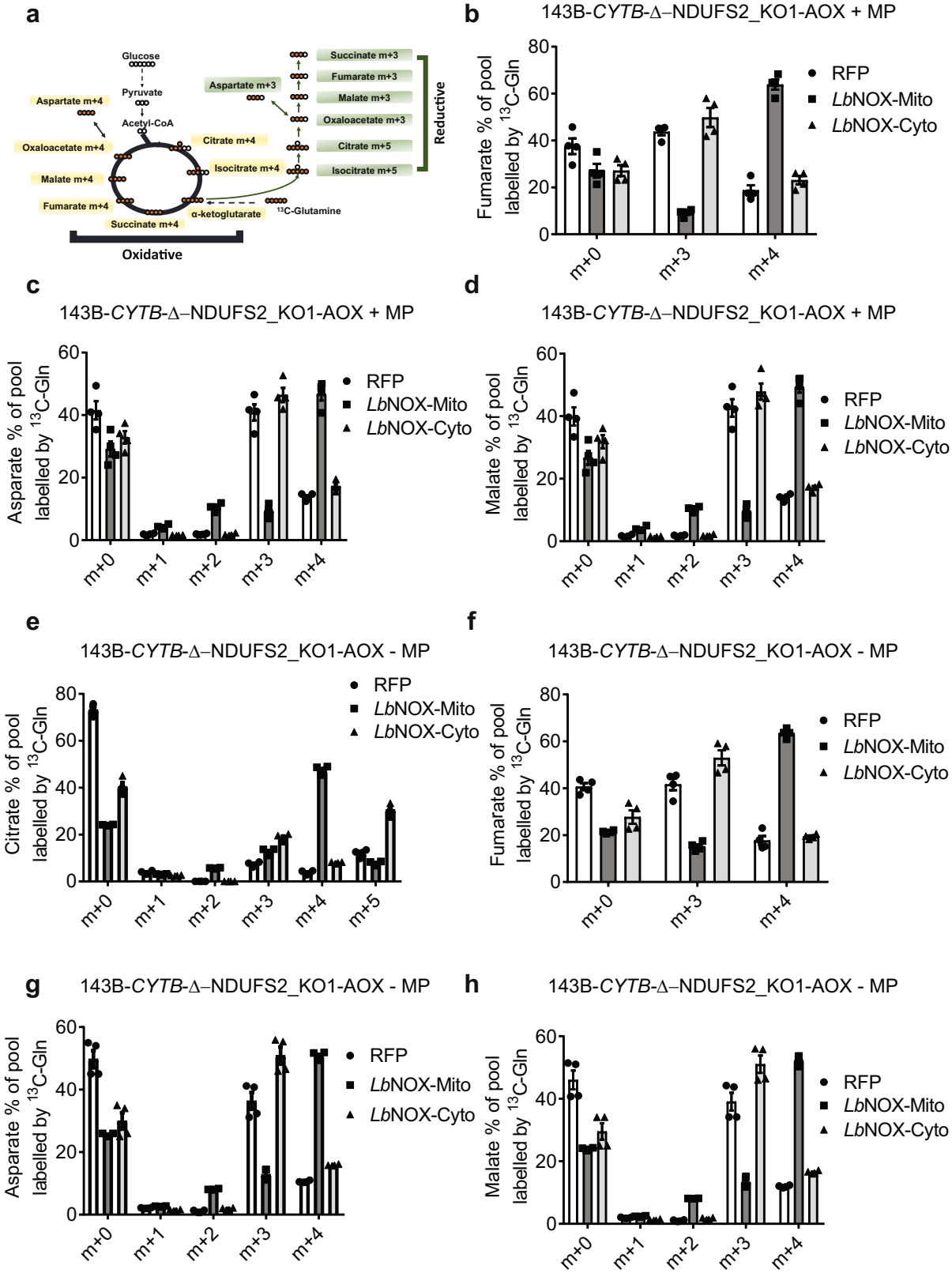
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | *Lb*NOX expression in mitochondria or cytosol promotes major changes in the metabolome of complex III-deficient cells. **a**, Schematic representation of the ETC in 143B-*CYTB*-Δ cells expressing *Lb*NOX in mitochondria. **b**, Intracellular NAD+/NADH ratio in 143B-*CYTB*-Δ-RFP, 143B-*CYTB*-Δ-*Lb*NOX-Mito and 143B-*CYTB*-Δ-*Lb*NOX-Cyto cells in the absence of methyl pyruvate ($n = 5$ biologically independent experiments). **c**–**e**, Intracellular aspartate (**c**), succinate (**d**) and α-ketoglutarate levels (**e**) in 143B-*CYTB*-Δ-RFP, 143B-*CYTB*-Δ-*Lb*NOX-Mito and 143B-*CYTB*-Δ-*Lb*NOX-Cyto cells in the absence of methyl pyruvate ($n = 5$ biologically independent experiments). **f**, The heat map displays the relative abundance of significantly changed metabolites in 143B-*CYTB*-Δ-RFP, 143B-*CYTB*-Δ-*Lb*NOX-Mito and 143B-*CYTB*-Δ-*Lb*NOX-Cyto cells in the absence of methyl pyruvate ($n = 5$ biologically independent experiments). A red–blue colour scale depicts the abundance of the metabolites (red: high, blue: low). **g**, Tumour mass of xenografts from 143B-*CYTB*-Δ-RFP, 143B-*CYTB*-Δ-*Lb*NOX-Mito and 143B-*CYTB*-Δ-*Lb*NOX-Cyto cells ($n = 9$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**b**–**e**, **g**). $^{*}P < 0.05$, $^{**}P < 0.01$, one-way ANOVA (**b**, **g**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). Metabolites levels (**c**–**f**) were analysed with multiple one-way ANOVA using an FDR of 0.1 and Fisher's least significant difference test post hoc analyses $Q = 10\%$. Each row was analysed individually ($^{*}Q < 0.1$; exact *Q* values in Source Data).

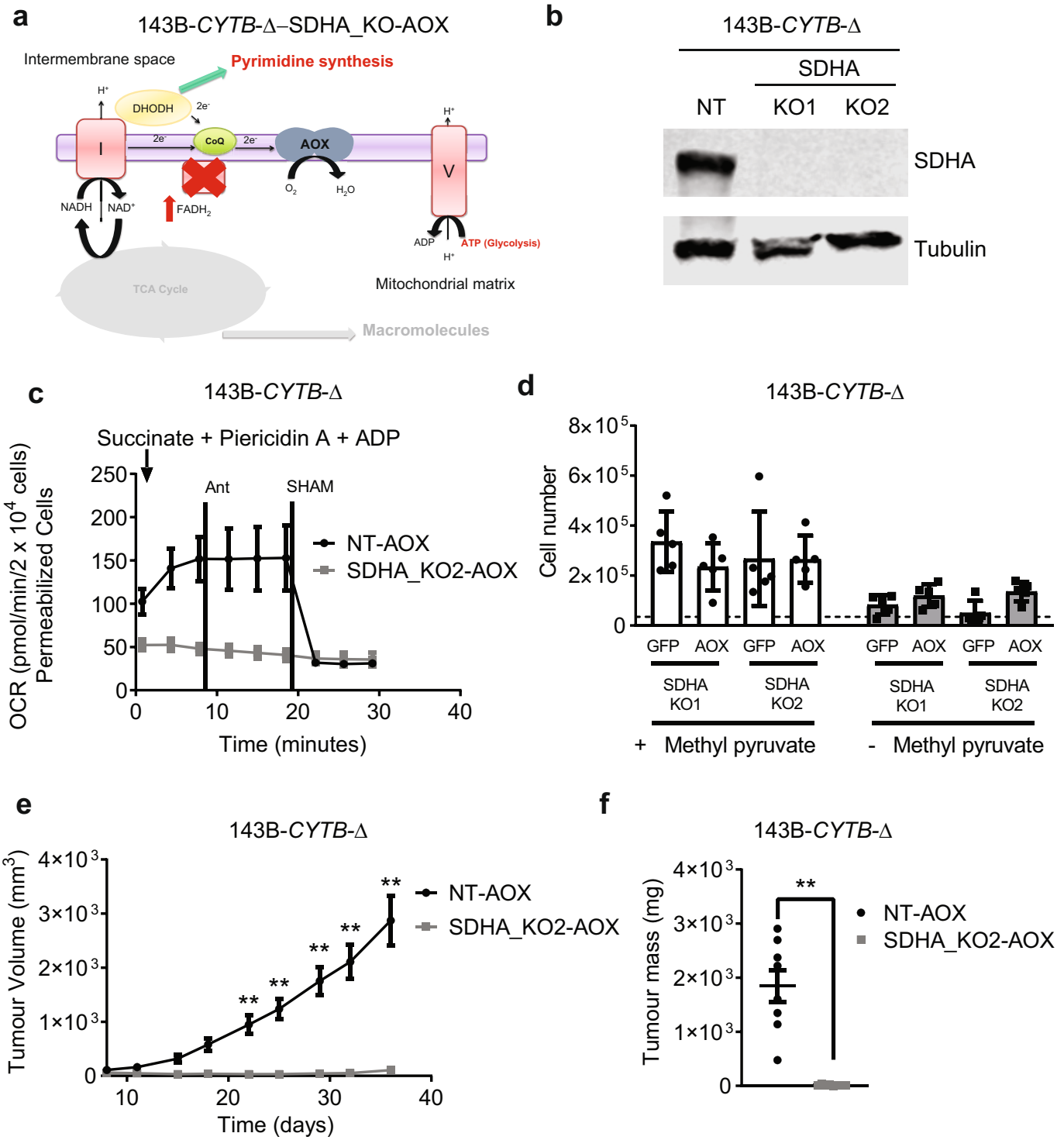


Extended Data Fig. 8 | See next page for caption.

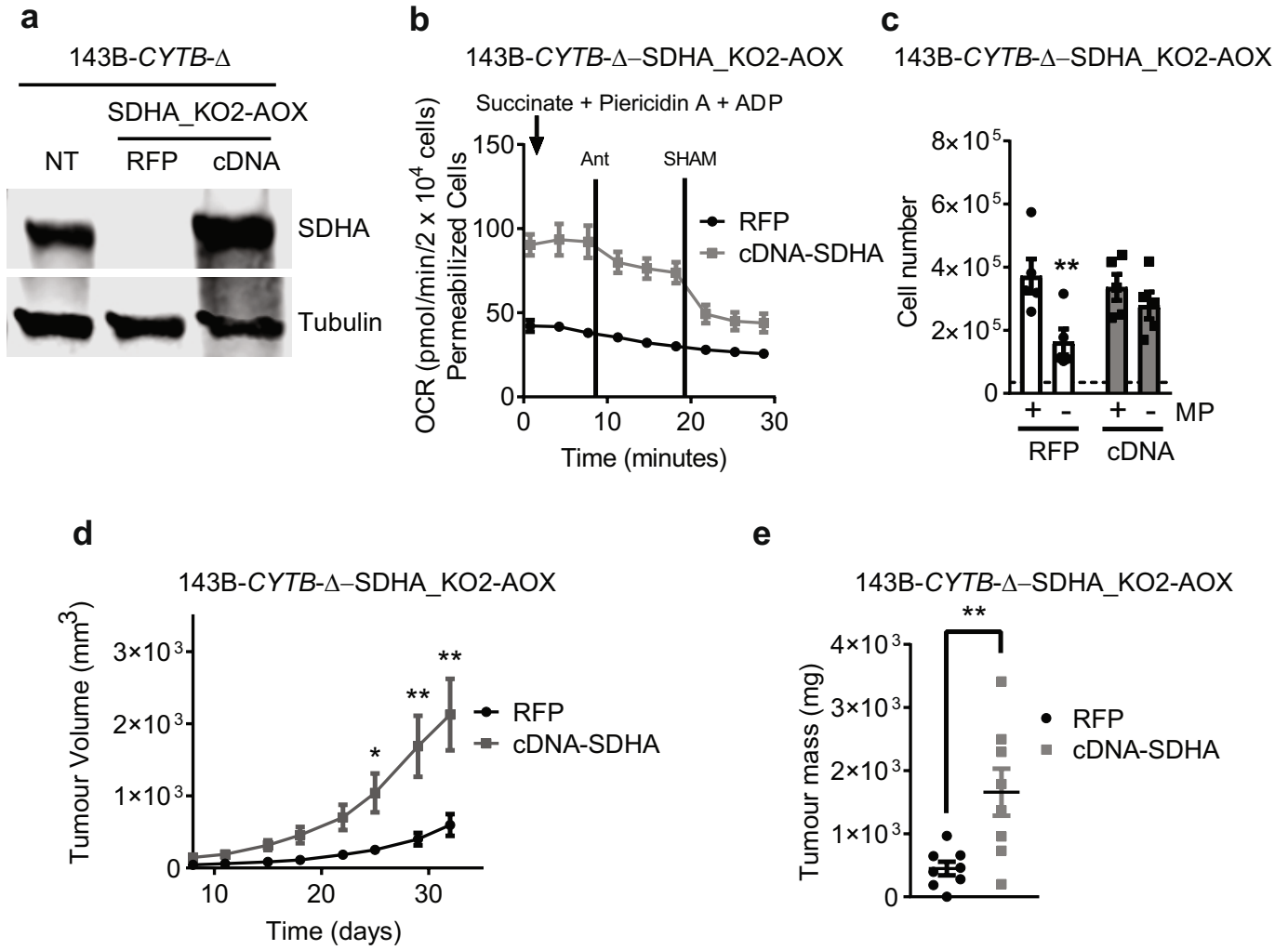
Extended Data Fig. 8 | *Lb*NOX expression in mitochondria or cytosol promotes major changes in the metabolome of complex I deficient cells. **a**, Schematic representation of the ETC in 143B-*CYTB*-Δ-NDUFS2_KO-AOX cells expressing *Lb*NOX in mitochondria. **b**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto were grown in the presence or absence of methyl pyruvate and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **c**, Intracellular NAD+/NADH ratio of 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto cells in the absence of methyl pyruvate and uridine ($n = 4$ biologically independent experiments). **d**, Intracellular aspartate levels of 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto cells in the absence of methyl pyruvate and uridine ($n = 4$ biologically independent experiments). **e**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto cells were grown in medium containing 10 mM glucose or 10 mM galactose for 48 h and assessed for cell death ($n = 4$ biologically independent experiments). **f**, Tumour mass of xenografts from 143B-*CYTB*-Δ-NDUFS2_KO1-AOX cells expressing *Lb*NOX in either mitochondria or cytosol ($n = 10$ mice per group from two independent cohorts). **g**, Western blot analysis (data representative of two independent experiments) of *Lb*NOX expression in xenograft tumours from 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto cells. Tubulin was used as a loading control. Data are mean ± s.e.m. (**b**–**f**). $^{*}P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**f**), one-way ANOVA (**c**) with a Bonferroni test for multiple comparisons or a two-way ANOVA (**b**, **e**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). Metabolites levels (**d**) were analysed with multiple one-way ANOVA using an FDR of 0.1 and Fisher's least significant difference test post hoc analyses $Q = 10\%$. Each row was analysed individually. ($^{*}Q < 0.1$; exact *Q* values in Source Data.) For gel source data, see Supplementary Fig. 5.



Extended Data Fig. 9 | Complex I-deficient cells expressing *Lb*NOX in the cytosol perform glutamine reductive carboxylation. a, Schematic representation for oxidative and reductive glutamine metabolism. Metabolism of [U-^{13}C]glutamine generates fully labelled α-ketoglutarate. Oxidation of α-ketoglutarate in the TCA cycle produces metabolites with four ^{13}C-carbons ($m+4$), while reduction of α-ketoglutarate through the reductive carboxylation pathway produces citrate with five ^{13}C-carbons ($m+5$). Further reductive metabolism of the $m+5$ citrate yields metabolites with three ^{13}C-carbons ($m+3$). **b**–**h**, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-RFP, 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Mito and 143B-*CYTB*-Δ-NDUFS2_KO1-AOX-*Lb*NOX-Cyto cells were labelled for 6 h with [U-^{13}C]glutamine in the presence (**b**–**d**) or absence (**e**–**h**) of methyl pyruvate, and the percentage of labelled metabolite pools was examined. $m+5$ and $m+3$ pools result from glutamine flow through reductive metabolism. $m+4$ pools result from glutamine flow through oxidative metabolism. Data are mean ± s.e.m. of four biologically independent experiments.



Extended Data Fig. 10 | Complex II is necessary for tumour growth. **a**, Schematic representation of the ETC in complex II deficient 143B-*CYTB*-Δ cells expressing AOX. **b**, Western blot analysis of SDHA in 143B-*CYTB*-Δ non-targeting and 143B-*CYTB*-Δ-SDHA_KO cells. Tubulin was used as a loading control. Data representative of two independent experiments. **c**, Complex II-driven OCR of permeabilized 143B-*CYTB*-Δ-NT-AOX and 143B-*CYTB*-Δ-SDHA_KO2-AOX cells. Piericidin A (1 µM) and antimycin A (1 µM) were used to inhibit complex I and III, respectively. SHAM (2 mM) was used to inhibit AOX activity ($n = 4$ biologically independent experiments). **d**, 143B-*CYTB*-Δ-SDHA-KOs expressing GFP or AOX were grown in the presence or absence of methyl pyruvate and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **e**, **f**, Average tumour volume (**e**) and tumour mass (**f**) of xenografts from 143B-*CYTB*-Δ-NT-AOX and 143B-*CYTB*-Δ-SDHA_KO2-AOX cells ($n = 8$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**c**–**f**). $^*P < 0.05$; $^{**}P < 0.01$, two-tailed *t*-tests (**f**) or two-way ANOVA (**d**, **e**) with a Bonferroni test for multiple comparisons (exact *P* value are in the Source Data). For gel source data, see Supplementary Fig. 6.



Extended Data Fig. 11 | Restoration of complex II by ectopic expression of *SDHA* cDNA rescues tumour growth. a, Western blot analysis of SDHA protein levels in 143B-*CYTB*-Δ-NT, 143B-*CYTB*-Δ-SDHA_KO2-AOX-RFP and 143B-*CYTB*-Δ-SDHA_KO2-AOX-cDNA SDHA cells. Data representative of three independent experiments. **b**, Complex II-driven OCR of permeabilized 143B-*CYTB*-Δ-SDHA_KO2-AOX-RFP and 143B-*CYTB*-Δ-SDHA_KO2-AOX-cDNA SDHA cells. Succinate and ADP were provided as substrates. Piericidin A (1 µM) and antimycin A (1 µM) were used to inhibit complex I and III respectively. SHAM (2 mM) was used to inhibit AOX activity ($n = 4$ biologically independent experiments). **c**, 143B-*CYTB*-Δ-SDHA_KO2-AOX-RFP and 143B-*CYTB*-Δ-SDHA_KO2-AOX-cDNA SDHA cells were grown in the presence or absence of methyl pyruvate and cell number was assessed after 72 h ($n = 5$ biologically independent experiments). **d**, **e**, Average tumour volume (**d**) and tumour mass (**e**) of xenografts from 143B-*CYTB*-Δ-SDHA_KO2-AOX-RFP and 143B-*CYTB*-Δ-SDHA_KO2-AOX-cDNA SDHA cells ($n = 8$ mice per group from two independent cohorts). Data are mean ± s.e.m. (**b**–**e**). $^*P < 0.05$, $^{**}P < 0.01$, two-tailed *t*-tests (**e**) or two-way ANOVA (**c**, **d**) with a Bonferroni test for multiple comparisons (exact *P* values are in the Source Data). For gel source data, see Supplementary Fig. 6.

nature research

Corresponding author(s): Navdeep S Chandel

Last updated by author(s): Mar 2, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Oxygen consumption data was collected using Wave 2.4 software. Flow cytometry data was collected using FACS DIVA 8.0.3 software. Metabolite data was collected using Xcalibur 4.1 software. Luminescence values from the tumors were collected using the IVIS 4.5 or the LAGO 2.3.1 imaging systems. Western blot images were collected using Odyssey Fc Imaging System 5.2 From LI-COR.
Data analysis	GraphPad Prism 7.0 and MetaboAnalyst 4.0 were used for statistical tests. Flow cytometry data was analyzed using Flowjo 10.4.2. Metabolite data was analyzed using Tracefinder 4.1 software. Images from tumors were processed using the Living Image 4.5 or the Aura 2.3.1 softwares. Image Studio Lite version 5.0 (LI-COR) was used for the analysis of protein levels.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data from the manuscript are available from the corresponding author on request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All the experiments were performed using sample sizes based on standard protocols in the field. We made an effort to avoid needless use of animals. No statistical test was performed to predetermine sample size. We used statistical analysis consistent with the sample size for each experiment and found sufficient statistical power with the sample sizes utilized in our study.
Data exclusions	For xenograft in vivo experiments, mice were excluded from the analysis if euthanasia had to be applied due to ulcerations or excessive tumor growth prior to the end point of the experiment. This exclusion criteria was pre-established.
Replication	All experimental data was reliably reproduced in multiple independent experiments as indicated in the figure legends. In vivo tumor experiments are from at least two independent cohorts to ensure reproducibility.
Randomization	Experimental animals were not randomized to experimental groups, but were age-matched, sex-matched, and littermates when possible.
Blinding	Investigators were not blinded. In case of leukemia studies blinding was not relevant, as groups consisted of previously genotyped mice in order to have correct experimental and control groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies used for flow cytometry were: anti-Mouse Ter-119 (eBioscience; #48-5921-80; clone TER-119), anti-Mouse NK1.1 (eBioscience; #48-5941-80; clone PK136,), anti-Human/Mouse CD45R (B220) (eBioscience; #48-0452-80; clone RA3-6B2), anti-Mouse CD8a (eBioscience; #48-0081-82; clone 53-6.7), anti-Mouse CD11b (eBioscience; #48-0112-80; clone M1/70,), anti-Mouse Ly-6G (Gr-1) (eBioscience; #48-5931-80; clone RB6-8C5,), anti-Mouse CD4 (Tonbo Biosciences; #75-0041-U100; clone GK1.5). Antibodies used for western blots were: anti-NDUFS2 (Abcam; #ab103024; 1:500 dilution), anti-QPC (Abcam; #ab136679; 1:500 dilution), anti-SDHA (MitoScience; #MS204; clone 2E3GC12FB2AE2; 1:500 dilution), anti-DHODH (Santa Cruz; #sc-166348; clone E-8; 1:500 dilution), anti-FLAG (Sigma; #F1804; clone M2; 1:1000 dilution), anti-GAPDH (Santa Cruz; #sc-32233; clone 6C5, and Sigma; #G9545; 1:2000 dilution), anti-ATP5A (Mitosciences; #MS507; clone 15H4C4; 1:1000 dilution), anti-Tubulin (Cell Signaling; #2144; 1:1000 dilution) and anti-β-Actin (Sigma; # A2228-100UL; clone AC-74,).
Validation	The antibodies used in this study were tested by the manufacturer. Antibodies used for flow citometry: - anti-Mouse Ter-119 (clone number: TER-119 eBioscience; catalogue number: 48-5921-80). This antibody has been used in 25 published figures and can be found in 42 references. The manufacturer also provides antibody testing data. https://www.thermofisher.com/antibody/product/TER-119-Antibody-clone-TER-119-Monoclonal/48-5921-80 - anti-Mouse NK1.1 (clone number: PK136, eBioscience; catalogue number: 48-5941-80). This antibody has been used in 27

published figures and can be found in 28 references. The manufacturer also provides antibody testing data. https://www.thermofisher.com/antibody/product/NK1-1-Antibody-clone-PK136-Monoclonal/48-5941-80

- anti-Human/Mouse CD45R (B220) (clone number: RA3-6B2, eBioscience; catalogue number: 48-0452-80). This antibody has been used in 40 published figures and can be found in 95 references. The manufacturer also provides antibody testing data and advanced verification by relative expression to ensure that the antibody binds to the antigen stated. https://www.thermofisher.com/antibody/product/CD45R-B220-Antibody-clone-RA3-6B2-Monoclonal/48-0452-80

- anti-Mouse CD8a (clone number: 53-6.7, eBioscience, catalogue number: 48-0081-82). This antibody has been used in 40 published figures and can be found in 138 references. The manufacturer also provides antibody testing data. https://www.thermofisher.com/antibody/product/CD8a-Antibody-clone-53-6-7-Monoclonal/48-0081-82

- anti-Mouse CD11b (clone number: M1/70, eBioscience; catalogue number: 48-0112-80). This antibody has been used in 40 published figures and can be found in 230 references. The manufacturer also provides antibody testing data. https://www.thermofisher.com/antibody/product/CD11b-Antibody-clone-M1-70-Monoclonal/48-0112-80

- anti-Mouse Ly-6G (Gr-1) (clone number: RB6-8C5, eBioscience; catalogue number: 48-5931-80). This antibody has been used in 40 published figures and can be found in 102 references. The manufacturer also provides antibody testing data and advanced verification by relative expression to ensure that the antibody binds to the antigen stated. https://www.thermofisher.com/antibody/product/Ly-6G-Ly-6C-Antibody-clone-RB6-8C5-Monoclonal/48-5931-80

- anti-Mouse CD4 (clone number: GK1.5, Tonbo Biosciences; catalogue number: 75-0041-U100). Tonbo Biosciences tests all antibodies by flow cytometry. https://tonbobio.com/products/violetfluor-450-anti-mouse-cd4-gk1-5

Antibodies used for western blots:

- anti-NDUFS2 (Abcam; catalogue number: ab103024; used at a 1:500 dilution). The Abpromise guarantee covers the used of the antibody for WB application. However, the antibody is not available anymore. https://www.abcam.com/ndufs2-antibody-ab103024.html

- anti-QPC (Abcam; catalogue number: ab136679; used at 1:500 dilution). The Abpromise guarantee covers the used of the antibody for WB application. https://www.abcam.com/uqcrq-antibody-ab136679.html#top-294

- anti-SDHA (clone number: 2E3GC12FB2AE2, MitoScience; catalog number: MS204; used at a 1:500 dilution). The Abpromise guarantee covers the used of the antibody for WB application. This antibody has been referenced in 235 publications. https://www.abcam.com/sdha-antibody-2e3gc12fb2ae2-ab14715.html#top-701

- anti-DHODH (clone number: E-8, Santa Cruz; catalog number: sc-166348; used at a 1:500 dilution). This antibody has been referenced in 5 publications. The manufacturer also provides antibody testing data. https://datasheets.scbt.com/sc-166348.pdf

- anti-FLAG (clone number: M2, Sigma; catalog number: F1804; used at a 1:1000 dilution). This antibody has been referenced in 4024 publications. The manufacturer also provides antibody testing data. https://www.sigmaaldrich.com/catalog/product/sigma/f1804?lang=en&region=US

- anti-GAPDH (clone number: 6C5, Santa Cruz; catalog number: sc-32233 and Sigma; catalog number: G9545; used at a 1:2000 dilution). For sc-32233 antibody from Santa Cruz, the antibody has been referenced in 2479 publications. The manufacturer also provides antibody testing data. https://datasheets.scbt.com/sc-32233.pdf. For G9545 from Sigma, the antibody has been referenced in 792 publications. The manufacturer also provides antibody testing data. https://www.sigmaaldrich.com/catalog/product/sigma/g9545?lang=en&region=US

- anti-ATP5A (clone number: 15H4C4, Mitosciences; catalog number: MS507; used at a 1:1000 dilution). The Abpromise guarantee covers the used of the antibody for WB application. The antibody has been referenced in 243 publications. The manufacturer also provides antibody testing data. https://www.abcam.com/atp5a-antibody-15h4c4-mitochondrial-marker-ab14748.html

- anti-Tubulin (Cell Signaling; catalog number: 2144, used at a 1:1000 dilution). This antibody has been referenced in 341 publications. The company tested that The α-Tubulin Antibody detects endogenous levels of total α-tubulin protein, and does not cross-react with recombinant β-tubulin. https://www.cellsignal.com/products/primary-antibodies/a-tubulin-antibody/2144

- anti-β-Actin (clone number: AC-74, Sigma; catalog number: A2228-100UL). This antibody has been referenced in 1412 publications. The manufacturer also provides antibody testing data. https://www.sigmaaldrich.com/catalog/product/sigma/a2228?lang=en&region=US

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	143B-Cytb-WT and 143B-Cytb-Δ cells were previously described (Weinberg et al. PNAS. 2010). Mouse KrasG12D p53-/- (KP) lung tumor cells expressing luciferase were generously provided by Dr. T. Papagiannakopoulos. 293T were from ATCC. Platinum E-retroviral packaging cells were a kind gift from Dr. P. Ntziachristos.
Authentication	Neither of the cell lines used were authenticated.

Mycoplasma contamination	Cell lines tested negative for mycoplasma contamination. Cells were checked periodically.
Commonly misidentified lines (See ICLAC register)	These cell lines are not listed in the database of commonly misidentified cell lines maintained by ICLAC.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	For leukemia experiments we used C57BL/6 Uqcrq (QPC null/wt or QPC null/fx) that have been recently described (Weinberg et al. Nature 2019). Uqcrq (QPC) floxed (Fx), wildtype (WT) and null (-) alleles were genotyped using the following primers: QPC-F: CTTCCGCTCCTCCCGGAAGT, QPC-R: TTCCCAAACTCGCGGCCCATG and QPC-null: CAATTCCAGCCAACAGTCCC. Ubc-CreERT2 mice were obtained from the Jackson Laboratory. Wild-type C57Bl/6 were used as recipients for T-ALL experiments. 8 to 12 weeks old mice of both sexes were used for experiments. For xenograft tumor studies, we used male J:Nu mice (8 to 12 weeks) obtained from Jackson Laboratory. For the orthotopic lung tumor model, we used male C57BL/6 mice (8 to 12 weeks).
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All mouse work was done in accordance with Northwestern University Institutional Animal Care and Use Committee (IACUC).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Sample preparation is described in detail in the methods section of the manuscript. Briefly, for T-ALL experiments, lineage (CD4, CD8a, B220, CD11b, Gr-1, NK1.1, Ter-119)-negative, and GFP+ bone marrow cells transduced with the Notch1ΔE-GFP retrovirus were sorted on BD FACS Aria systems. To asses tumor burden, spleen, and one set of pelvis, femur, and tibia were harvested from each recipient, and analyzed for the number and % of GFP+ T-ALL cells, using PKH reference microbeads (Sigma). Samples were harvested from adult mice. To obtain a single-cell suspension, tissues were disrupted using scored 60mm petri dishes in PBS containing 2% FBS and filtered through a 70µM nylon mesh filter. Expression of GFP was analyzed using BD FACSymphony. For in vitro proliferation experiments, supernatant containing dead cells and cells attached to the wells were collected and centrifuged at 300 × g for 5 min. AccuCount Fluorescent Particles from Spherotech were added to count the cells. Cell viability was assessed by DAPI staining. Numerical values for number of cells or percentage with statistics for each graph is provided in Source Data files.
Instrument	BD LSR Fortessa, BD FACSymphony or BD FACSAria cell sorter.
Software	BD FASC Diva was used for collection of the data. All data was analyzed using FlowJo software.
Cell population abundance	The cells were periodically sorted to maintain high protein expressions. An aliquot of the sorted cells were always collected and run on a cytometer to verify purity of the samples collected. In addition, cell counts were performed on samples post sort to verify correct cell numbers.
Gating strategy	A supplemental figure is provided to show the gating strategy for T-ALL burden analysis. 2 different analysis were done: # of GFP+ Cells: Used the following flow plots to obtain: SSC-A vs. FSC-A: count of Lymphocytes --> A FSC-H vs. PE-A: count of PKH beads --> B APC-Cy7-A vs. BB515-A (showing the Lymphocytes population): % of GFP+ and Live lymphocytes --> C Actual # of lymphocytes = (A / B) x PKH bead concentration (in beads/ml, counted using hemocytometer each time) x total volume of cell suspension (ml) # of GFP+ cells = Actual # of lymphocytes x C

% of GFP+ Cells:
SSC-A vs. FSC-A: Lymphocytes population
FSC-H vs. FSC-A (showing the Lymphocytes population): Singlet population
SSC-H vs. APC-Cy7-A (showing the Singlet population): Live population
SSC-H vs. BB515-A (showing the live population): % of GFP+ population --> D
% GFP+ Cells = D

For proliferation experiments, beads and cells were identified first by discrimination by size (FSC-A by SSC-A). Singlets were then distinguished using FSC-A by FSH-H. Beads and Live cells were further gated by the appropiate fluorochrome used.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

The CDK inhibitor CR8 acts as a molecular glue degrader that depletes cyclin K

https://doi.org/10.1038/s41586-020-2374-x

Received: 9 November 2019

Accepted: 29 April 2020

Published online: 3 June 2020

Check for updates

Mikołaj Słabicki[1,2,3,15], Zuzanna Kozicka[4,5,15], Georg Petzold[4,15], Yen-Der Li[1,2,6], Manisha Manojkumar[1,2,3], Richard D. Bunker[4,14], Katherine A. Donovan[7,8], Quinlan L. Sievers[1,2], Jonas Koeppel[1,2,3], Dakota Suchyta[4,5], Adam S. Sperling[1,2], Emma C. Fink[1,2], Jessica A. Gasser[1,2], Li R. Wang[1], Steven M. Corsello[1,2], Rob S. Sellar[1,2,9], Max Jan[1,2], Dennis Gillingham[5], Claudia Scholl[10], Stefan Fröhling[3,11], Todd R. Golub[1,12,13], Eric S. Fischer[7,8], Nicolas H. Thomä[4]✉ & Benjamin L. Ebert[1,2,13]✉

Molecular glue compounds induce protein–protein interactions that, in the context of a ubiquitin ligase, lead to protein degradation[1]. Unlike traditional enzyme inhibitors, these molecular glue degraders act substoichiometrically to catalyse the rapid depletion of previously inaccessible targets[2]. They are clinically effective and highly sought-after, but have thus far only been discovered serendipitously. Here, through systematically mining databases for correlations between the cytotoxicity of 4,518 clinical and preclinical small molecules and the expression levels of E3 ligase components across hundreds of human cancer cell lines[3–5], we identify CR8—a cyclin-dependent kinase (CDK) inhibitor[6]—as a compound that acts as a molecular glue degrader. The CDK-bound form of CR8 has a solvent-exposed pyridyl moiety that induces the formation of a complex between CDK12–cyclin K and the CUL4 adaptor protein DDB1, bypassing the requirement for a substrate receptor and presenting cyclin K for ubiquitination and degradation. Our studies demonstrate that chemical alteration of surface-exposed moieties can confer gain-of-function glue properties to an inhibitor, and we propose this as a broader strategy through which target-binding molecules could be converted into molecular glues.

Molecular glues are a class of small-molecule drugs that induce or stabilize interactions between proteins[1]. In the context of a ubiquitin ligase, drug-induced interactions can lead to protein degradation, and this is an emerging strategy for the inactivation of therapeutic targets that are intractable by conventional pharmacological means[2]. Known molecular glue degraders bind to the substrate receptors of E3 ubiquitin ligases and recruit target proteins for their ubiquitination and subsequent degradation by the proteasome.

Thalidomide analogues and aryl sulfonamides are two classes of drugs that act as molecular glue degraders. Widely used in the clinic, thalidomide analogues are an effective treatment for multiple myeloma, other B cell malignancies and myelodysplastic syndrome with a deletion in chromosome 5q[7]. Thalidomide analogues recruit zinc-finger transcription factors and other target proteins to cereblon (CRBN)[8–11], the substrate receptor of the cullin-RING E3 ubiquitin ligase CUL4A/B–RBX1–DDB1–CRBN ($CRL4^{CRBN}$)[12]. Similarly, aryl sulfonamides degrade the essential RNA-binding protein RBM39 by engaging DCAF15, the substrate receptor of the $CRL4^{DCAF15}$ E3 ubiquitin ligase[13–15]. In these examples, the degraders are not dependent on a ligandable pocket on the target protein, but instead exploit complementary protein–protein interfaces between the receptor and the target. By reprogramming the selectivity of the ubiquitin ligase, these molecules divert the ligase to drive multiple rounds of target ubiquitination in a catalytic manner[16]. Such compounds can thus circumvent the limitations of classical inhibitors and expand the repertoire of 'druggable' proteins. Although highly desirable, molecular glue degraders have only been found serendipitously, and the strategies available for identifying or designing these compounds are limited.

CR8 induces proteasomal cyclin K degradation

To identify small molecules that mediate protein degradation through an E3 ubiquitin ligase, we correlated drug-sensitivity data for 4,518 clinical and preclinical drugs that have been tested against 578 cancer cell lines[3,4] with the respective mRNA levels of 499 E3 ligase components[5] (Extended Data Fig. 1a). Expression of *DCAF15* correlated with

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [3]Division of Translational Medical Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [4]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [5]Faculty of Science, University of Basel, Basel, Switzerland. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [8]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Haematology, UCL Cancer Institute, University College London, London, UK. [10]Division of Applied Functional Genomics, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [11]German Cancer Consortium, Heidelberg, Germany. [12]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [13]Howard Hughes Medical Institute, Boston, MA, USA. [14]Present address: Monte Rosa Therapeutics, Basel, Switzerland. [15]These authors contributed equally: Mikołaj Słabicki, Zuzanna Kozicka, Georg Petzold. ✉e-mail: nicolas.thoma@fmi.ch; benjamin_ebert@dfci.harvard.edu

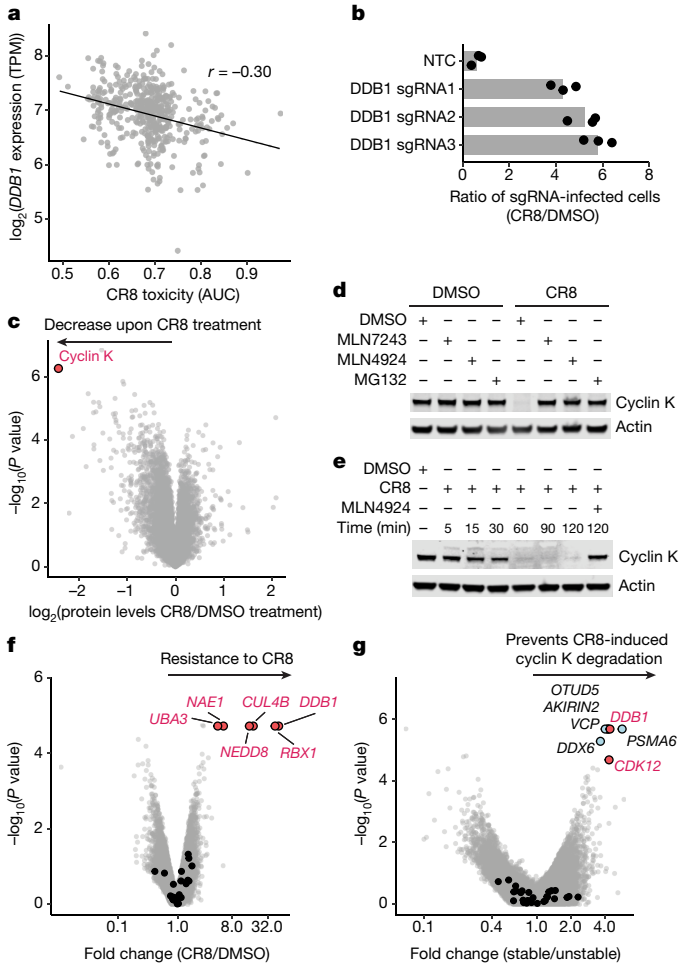


Fig. 1 | CR8-induced degradation of cyclin K depends on DDB1 and CDK12. **a**, Pearson correlation between CR8 toxicity and *DDB1* mRNA levels. Dots represent cancer cell lines ($n = 471$). A lower value for the area under the curve (AUC; *y* axis) corresponds to higher drug toxicity. The Pearson correlation coefficient (r) is shown. TPM, transcripts per million. **b**, Flow cytometry analysis of HEK293T-Cas9 cells expressing three different sgRNAs against *DDB1* and a blue fluorescent protein (BFP) marker after a three-day treatment with 1 μM CR8 (bars represent mean, $n = 3$). NTC, non-targeting control. **c**, Whole-proteome quantification of MOLT-4 cells treated with 1 μM CR8 ($n = 1$) or DMSO ($n = 3$) for 5 h (two-sided moderated *t*-test, $n = 3$). **d**, Immunoblots of cyclin K degradation in HEK293T-Cas9 cells that were pretreated with 0.5 μM MLN7243, 1 μM MLN4924 or 10 μM MG132 for 4 h and then treated with 1 μM CR8 for 2 h ($n = 3$). **e**, Immunoblots of the time course of cyclin K degradation in HEK293T-Cas9 cells treated with 1 μM CR8 ($n = 3$). **f**, Genome-wide CRISPR–Cas9 resistance screen for CR8 resistance in HEK293T-Cas9 cells. **g**, Genome-wide CRISPR–Cas9 reporter screen for cyclin K–eGFP stability after treatment with 1 μM CR8 in HEK293T-Cas9 cells. In **f**, **g**, guide counts were collapsed to gene level ($n = 4$ guides per gene; two-sided empirical rank-sum test statistics). Black dots denote DCAF substrate receptors (**f**, **g**).

indisulam and tasisulam toxicity, consistent with the known function of these drugs as degraders of the essential protein RBM39 through the $CRL4^{DCAF15}$ E3 ubiquitin ligase and thus demonstrating the potential of the approach (Extended Data Fig. 1b, c). We sought to validate the correlations between ligase expression and drug toxicity that scored most highly by examining whether CRISPR-mediated inactivation of the identified E3 ligase component would rescue the respective drug-induced toxicity (Extended Data Fig. 1d). These experiments confirmed that single-guide RNAs (sgRNAs) that target *DCAF15* confer resistance to indisulam and tasisulam. In addition, we observed a correlation between the cytotoxicity of (*R*)-CR8, a CDK inhibitor[6], and the mRNA levels of the CUL4 adaptor protein DDB1 (Fig. 1a, Extended Data Fig. 1e). Consistently, sgRNAs targeting *DDB1* conferred resistance to (*R*)-CR8 (Fig. 1b).

As the DDB1-dependent cytotoxicity of (*R*)-CR8 suggested ubiquitin ligase-mediated degradation of one or more essential proteins, we performed quantitative proteome-wide mass spectrometry to evaluate protein abundance after treating cells with (*R*)-CR8. Of the quantified proteins (more than 8,000) cyclin K was the only protein that consistently showed a decrease in abundance after addition of (*R*)-CR8 (Fig. 1c, Extended Data Fig. 1f, g). As expected, (*R*)-CR8 did not alter the levels of cyclin K mRNA (Extended Data Fig. 1h) and the (*R*)-CR8-induced degradation of cyclin K could be rescued by inhibition of the E1 ubiquitin-activating enzyme (using MLN7243), inhibition of cullin neddylation (MLN4924) and inhibition of the proteasome (MG132) (Fig. 1d). Together, these results suggest that (*R*)-CR8 triggers rapid proteasomal degradation of cyclin K (Fig. 1e) through the activity of a DDB1-containing cullin-RING ubiquitin ligase.

To dissect the molecular machinery that is required for (*R*)-CR8 toxicity, we performed genome-wide and E3 ubiquitin ligase-focused CRISPR–Cas9 resistance screens (Fig. 1f, Extended Data Fig. 2a, b). sgRNAs that target *DDB1*, *CUL4B*, *RBX1*, the cullin-RING activator *NEDD8* and the NEDD8-activating enzyme (*NAE1* and *UBA3*) were substiantially enriched in the (*R*)-CR8-resistant cell population. As all of these proteins are required for the activity of cullin-RING ligases, our results provide genetic evidence for the involvement of a functional CUL4–RBX1–DDB1 ubiquitin ligase complex in mediating (*R*)-CR8 cytotoxicity.

Thus far, all known cullin-RING ligases engage their substrates through specific substrate receptors, and DDB1 serves as an adaptor protein that is able to bind over 20 such receptors (also known as DDB1–CUL4-associated-factors, DCAFs)[17,18] to recruit them to the CUL4–RBX1 ligase core. As no DCAFs were identified in our (R)-CR8 resistance screens, we constructed a fluorescent reporter of cyclin K stability (Extended Data Fig. 2c), in which the (*R*)-CR8-mediated degradation of endogenous cyclin K (Fig. 1d, e) could be recapitulated with cyclin K fused to enhanced green fluorescent protein (eGFP) (cyclin K_{eGFP}) (Extended Data Fig. 2d–f). Using the stability reporter, in which the extent of degradation can be determined by measuring the levels of cyclin K_{eGFP} normalized to mCherry expression, we found that both (*S*)-CR8 and (*R*)-CR8 facilitated the degradation of cyclin K_{eGFP} to the same extent (Extended Data Fig. 2g; henceforth, CR8 refers to (*R*)-CR8). We then performed a genome-wide CRISPR–Cas9 screen for genes involved in cyclin K reporter stability and validated the involvement of DDB1 in CR8-mediated, but not CR8-independent, degradation of cyclin K (Fig. 1g, Extended Data Fig. 2h–j). In addition, we identified CDK12—a known target of CR8[19] that depends on the interaction with cyclin K for its activity[20]—as a crucial component for CR8-induced destabilization of cyclin K_{eGFP} (Fig. 1g, Extended Data Fig. 2h–k).

As neither the cyclin K_{eGFP} stability reporter screen nor the CR8 resistance screen identified a substrate receptor, we performed additional CRISPR screens targeting 29 genes that encode known DCAFs or DCAF-like candidate proteins in four different cell lines. Although sgRNAs targeting the previously identified components of the CUL4–RBX1–DDB1 complex consistently caused resistance to CR8, a DCAF substrate receptor could not be identified (Extended Data Fig. 3).

CR8 directs CDK12 to the ligase core

As none of our genetic screens identified a DCAF that is required for cyclin K degradation, we tested whether the CR8-engaged CDK12–cyclin K complex directly binds one of the CUL4–RBX1–DDB1 ligase components in the absence of a substrate receptor. We therefore performed in vitro co-immunoprecipitation experiments using recombinantly purified proteins. The kinase domain of CDK12

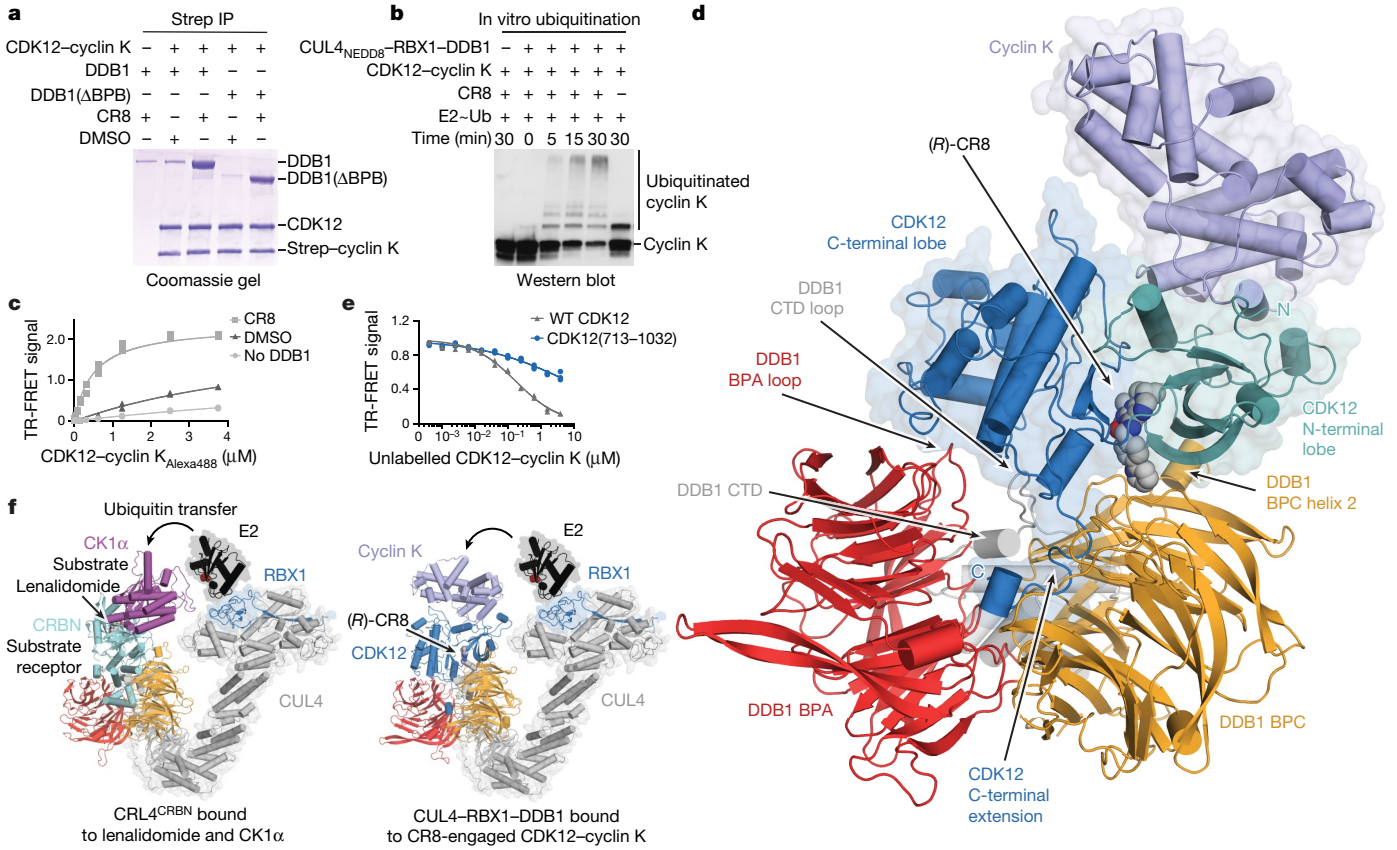


Fig. 2 | CR8-bound CDK12 binds DDB1 in a DCAF-like manner. **a**, Co-immunoprecipitation (IP) experiments with recombinant proteins ($n = 3$). Strep, streptavidin. **b**, In vitro ubiquitination of cyclin K by the $CUL4_{NEDD8}$-RBX1–DDB1 ubiquitin ligase core ($n = 2$). In E2-Ub, ~ represents the thioester bond between the active site of the E2 and the C terminus of ubiquitin. **c**, TR-FRET signal for CDK12–cyclin $K_{Alexa488}$ titrated to DDB1 tagged with terbium-coupled streptavidin ($DDB1_{terbium}$) in DMSO or 10 μM CR8 ($n = 3$). 'No DDB1' only contains terbium-coupled streptavidin and shows concentration-dependent fluorophore effects. **d**, Cartoon representation of the crystal structure of DDB1(ΔBPB)–(*R*)-CR8–CDK12–cyclin K. CTD, C-terminal domain. **e**, TR-FRET counter-titration of unlabelled wild-type (WT) or mutant CDK12–cyclin K (0–10 μM) into the preassembled $DDB1_{terbium}$–CR8–CDK12–cyclin $K_{Alexa488}$ complex ($n = 3$). **f**, Structural models of $CRL4^{CRBN}$ bound to lenalidomide and CK1α (left) and CUL4–RBX1–DDB1 (CRL4) bound to the (*R*)-CR8–CDK12–cyclin K complex (right). A cysteine residue at the active site of the E2 enzyme (red spheres) binds ubiquitin through a thioester bond.

(CDK12(713–1052)) bound to cyclin K(1–267) did not markedly enrich DDB1 over the bead-binding control in the absence of CR8, whereas equimolar amounts of CR8 led to stoichiometric complex formation (Fig. 2a). The DDB1 β-propeller domain A (BPA) and β-propeller domain C (BPC)[17], which are otherwise involved in DCAF binding, were sufficient for drug-induced recruitment of CDK12–cyclin K. DDB1 β-propeller domain B (BPB), which binds CUL4 and is not involved in DCAF binding, was dispensable for the interaction (Fig. 2a). In vitro ubiquitination assays confirmed that the CUL4A–RBX1–DDB1 ligase core alone is sufficient to drive robust ubiquitination of cyclin K (Fig. 2b). Quantification of the interaction showed that CR8 stimulated binding between CDK12–cyclin K and DDB1 in the range of 100–500 nM, depending on the experimental set-up (Fig. 2c, Extended Data Fig. 4). Although a weak interaction between CDK12–cyclin K and DDB1 was still detectable in vitro in the absence of the drug, CR8 strengthened complex formation by 500- to 1,000-fold as estimated by isothermal titration calorimetry (ITC) (Extended Data Fig. 4f–k). Thus, our data indicate that CR8-engaged CDK12–cyclin K is recruited to the CUL4–RBX1–DDB1 ligase core through DDB1, and CR8 tightens the complex sufficiently to drive CR8-induced degradation of cyclin K in the absence of a canonical DCAF substrate receptor.

We then crystallized CDK12(713–1052)–cyclin K(1–267) bound to CR8 and a truncated version of DDB1 that lacks the BPB domain (ΔBPB), and determined the structure of this complex at 3.5 Å resolution (Fig. 2d, Extended Data Table 1). In the structure, CDK12 forms extensive protein–protein interactions (around 2,100 $Å^2$) with DDB1. CR8 binds the active site of CDK12 and bridges the CDK12–DDB1 interface, whereas cyclin K binds CDK12 on the opposite side and does not contact DDB1. The N-terminal and C-terminal lobes of CDK12 are proximal to DDB1 residues located in a loop of the BPA domain (amino acids 111–114), helix 2 of the BPC domain (amino acids 986–990) and a loop in the C-terminal domain (amino acids 1078–1081), which are otherwise involved in DCAF binding (Extended Data Fig. 5). In addition, the C-terminal extension of CDK12 binds the cleft between the DDB1 domains BPA and BPC—a hallmark binding site for interactions between DDB1 and DCAFs (Extended Data Fig. 5a–d, i). The density for this region could only be tentatively assigned, probably owing to the presence of multiple conformations, but the CDK12 C-terminal tail clearly engages with DDB1 and assumes a conformation that is different from those seen in isolated CDK12–cyclin K structures[19,21] (Extended Data Fig. 6a, b, d). Structure-guided mutational analyses combined with time-resolved fluorescence resonance energy transfer (TR-FRET) assays were used to assess the contribution of these interactions to the CR8-dependent formation of the CDK12–DDB1 complex (Fig. 2e, Extended Data Fig. 5e). Together, our data show that CDK12 assumes the role of a glue-induced substrate receptor and places cyclin K in a position that is typically occupied by CRL4 substrates (Fig. 2f). This renders the binding of CDK12–cyclin K to DDB1 mutually exclusive with that of DCAFs and provides a structural framework that explains why a canonical substrate receptor is dispensable for cyclin K ubiquitination.

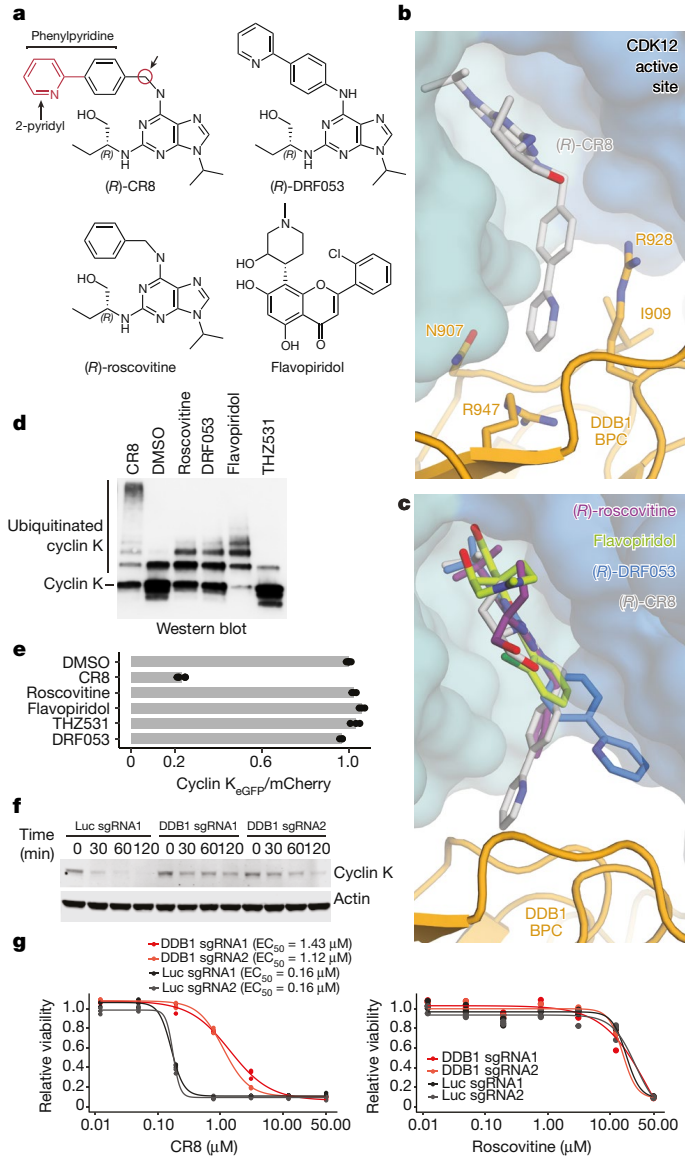


Fig. 3 | A surface-exposed 2-pyridyl moiety of CR8 confers glue degrader activity. a, Chemical structures of CDK inhibitors. Arrows indicate differences between (*R*)-CR8, (*R*)-DRF053 and (*R*)-roscovitine. **b**, Close-up view of the DDB1–CR8–CDK12 interface. The phenylpyridine moiety of CR8 contacts DDB1 residues. **c**, (*R*)-roscovitine (Protein Data Bank (PDB) entry 2A4L), (*R*)-DRF053 and flavopiridol (3BLR) in the active site of CDK12 in the DDB1–CR8–CDK12–cyclin K complex through superposition of kinase domains or the purine moiety (for DRF053). **d**, In vitro ubiquitination of the CDK12–cyclin K complex by $CUL4_{NEDD8}$–RBX1–DDB1 in the absence (DMSO) or presence of the indicated compounds (all 2 μM) ($n = 2$). **e**, Flow cytometry analysis of the degradation of cyclin K_{eGFP} in HEK293T-Cas9 cells treated with 1 μM of the indicated compounds for 2 h ($n = 3$). **f**, Immunoblots of cyclin K in HEK293T-Cas9 cells transfected with the indicated sgRNAs and treated with 1 μM CR8 ($n = 2$). **g**, Drug sensitivity of sgRNA-transfected HEK293T-Cas9 cells after three days of treatment with CR8 (left) or roscovitine (right) ($n = 3$). The half-maximum effective concentration (EC_{50}) values are shown for cells treated with CR8.

The CDK12–DDB1 interface imparts selectivity

CR8 is a pleiotropic CDK inhibitor that is reported to bind CDK1, CDK2, CDK3, CDK5, CDK7, CDK9 and CDK12[6,19], yet in cells we observed selective destabilization of cyclin K in the presence of the drug. As cyclin K is reported to associate with CDK9, CDK12 and CDK13[19], we tested whether the other cyclin K-dependent kinases are also recruited to DDB1. The closely related CDK13 (90.8% sequence identity)—but not the more divergent CDK9 (45.5% sequence identity) (Extended Data Fig. 7a–c)—was recruited to DDB1 in the presence of CR8, albeit with a lower binding affinity (Extended Data Fig. 7d–f). Analogously, in vitro ubiquitination of cyclin K was less productive for CDK13 than CDK12 (Extended Data Fig. 7g). The key difference in primary sequence between CDK9 and CDK12 or CDK13 is in the C-terminal extension (Extended Data Fig. 7a, b), which in our structure nestles up against the BPA and BPC regions of DDB1 (Fig. 2d, Extended Data Fig. 5i). Mutations in, or truncation of, the C-terminal extension of CDK12 abolished basal binding between CDK12 and DDB1, whereas complex formation could still be facilitated by CR8 to a varying extent (Extended Data Fig. 7h, i). Hence, our data show that the pan-selective CDK inhibitor CR8 induces specific protein–protein interactions between CDK12 or CDK13 and DDB1 and suggest that the C-terminal extension, though contributing to binding, is not essential for drug-dependent kinase recruitment.

CR8 phenylpyridine confers glue activity

CR8 occupies the ATP-binding pocket of CDK12 and forms discrete contacts with residues in the BPC domain of DDB1 (around 150 $Å^2$) through its hydrophobic phenylpyridine ring system (Fig. 3a, b). Mutations of the DDB1 residues Ile909, Arg928 and Arg947 each diminished drug-induced recruitment of the kinase (Extended Data Fig. 5f), highlighting the contribution of the phenylpyridine moiety to complex formation. To evaluate the structure–activity relationship that underlies the gain-of-function activity of CR8, we probed other CDK inhibitors for their ability to drive complex formation between DDB1 and CDK12. DRF053[22], an inhibitor related to CR8 that carries a differently linked phenylpyridine ring system (Fig. 3a, c), induced binding with a twofold lower affinity than CR8 (Extended Data Fig. 8a). Roscovitine[23], the parent compound of CR8, which lacks the 2-pyridyl substituent but retains the phenyl ring proximal to Arg928 (Fig. 3a, c), also facilitated complex formation—albeit with an apparent binding affinity threefold lower than CR8 (Extended Data Fig. 8a). The rank order of binding affinity that we observed in our TR-FRET assay correlated with the degree of cyclin K ubiquitination in vitro; DRF053 and roscovitine showed less processive ubiquitination (Fig. 3d). As neither DRF053 nor roscovitine induced degradation of the cyclin K_{eGFP} reporter in cells (Fig. 3e), our results show that the presence and correct orientation of the 2-pyridyl moiety on the surface of CDK12 confer the gain-of-function activity of CR8 that leads to cyclin K degradation.

To investigate whether any bound ligand could in principle drive the interaction of CDK12 with DDB1, we tested the endogenous CDK nucleotide cofactor ATP for its ability to promote complex formation. ATP neither facilitated nor abrogated the interaction over basal binding observed in the presence of dimethyl sulfoxide (DMSO) (Extended Data Fig. 6c), suggesting that although the nucleotide-bound conformation of CDK12 seems incompatible with the recruitment of DDB1 (Extended Data Fig. 6b), its C-terminal extension is free to adopt multiple conformations[21]. THZ531[24], a bulky covalent inhibitor of CDK12 and CDK13 that is predicted to clash with DDB1 (Extended Data Fig. 6d–f), locks the CDK12 C-terminal extension in a conformation that is incompatible with DDB1 recruitment (Extended Data Fig. 6d). Consistently, THZ531 further decreased the TR-FRET signal and diminished cyclin K ubiquitination in vitro below the levels of the DMSO control[24] (Fig. 3d, Extended Data Fig. 6c). Flavopiridol[25]—an inhibitor that is derived from a natural product and is structurally distinct from CR8 (Fig. 3a, c)—also stimulated the binding of CDK12–cyclin K to DDB1 (Extended Data Fig. 8a). Although flavopiridol led to moderate ubiquitination of cyclin K in vitro (Fig. 3d), it did not degrade cyclin K in cells (Fig. 3e). Our results thus show that the interactions between DDB1 and different inhibitor compounds display substantial plasticity and that structurally diverse surface-exposed moieties in CR8, DRF053, roscovitine and flavopiridol can facilitate CDK12–cyclin K recruitment.

Small differences in their ability to stabilize the DDB1–CDK12 complex translate, in an almost binary manner, into the cellular degradation of cyclin K or lack thereof. This behaviour is reminiscent of $CRL4^{CRBN}$ and thalidomide analogues[11,26], in which an apparent affinity threshold must be overcome to drive drug-induced degradation of the target protein.

Cyclin K degradation adds to CR8 toxicity

Finally, to examine how CRL4-mediated degradation of cyclin K contributes to CR8 cytotoxicity compared to non-degradative CDK inhibition, we analysed CR8 toxicity in wild-type HEK293T-Cas9 cells and cells that were pretreated with MLN4924 (a NEDD8-activating enzyme inhibitor), subject to DCAF overexpression or genetically depleted of DDB1. Global inhibition of the activity of cullin-RING ligases by MLN4924 had only minor effects on cell viability (Extended Data Fig. 9a), but resulted in decreased sensitivity to CR8 (Extended Data Fig. 9b), showing that the neddylation of cullin-RING ligases substantially contributes to CR8 toxicity. Overexpression of the substrate receptor CRBN also affected the sensitivity of cells to CR8 and decreased the degradation of cyclin K (Extended Data Fig. 9c–g), presumably by reducing the free pool of DDB1. As expected, CR8-induced degradation of cyclin K was dependent on DDB1 (Fig. 3f) and, consistently, we found that cytotoxicity of CR8—but not that of the other CDK inhibitors—was tenfold lower in cells depleted of DDB1 (Fig. 3g, Extended Data Fig. 9h). Together, the data demonstrate that the CRL4-dependent gain-of-function glue degrader activity of CR8 strongly contributes to its cellular potency and provides an additional layer of orthologue-specific CDK inactivation through cyclin K degradation.

Kinase inhibitors have long been suspected to have a degradation component to their mode of action[27,28], and our work provides a characterization and structural dissection of how a kinase inhibitor scaffold acquires degrader properties. Molecular glue degraders have thus far only been shown to engage substrate-recruiting E3 ligase modules. CDK12 is not a constitutive E3 ligase component, but rather serves as a drug-induced substrate receptor, linking DDB1 to the ubiquitination target. CR8 thus bypasses the requirement for a canonical DCAF and instead hijacks the essential adaptor protein DDB1. Although cyclin K is the primary ubiquitination target, CDK12 may become subject to autoubiquitination after prolonged exposure to CR8, in a similar manner to canonical DCAFs[29,30].

Whereas previously reported molecular glue degraders engage a ligandable pocket on the ligase to recruit target proteins, CR8 instead binds the ATP pocket of CDK12 and does not rely on an independent ligand-binding site on DDB1 (Extended Data Fig. 4h). This suggests that the repertoire of target proteins and ubiquitin ligases accessible to targeted degradation can be expanded through target-binding small molecules that induce de novo contacts with a ligase or strengthen existing weak protein–protein interactions. Kinase inhibitors in particular often show poor selectivity, and small-molecule-induced inactivation of kinases that leverages complementary protein–protein interfaces offers a path towards improved drug selectivity—which might, for example, facilitate the selective inactivation of CDK12, an emerging therapeutic target[31].

The gain-of-function glue degrader activity of CR8 is attributed to a 2-pyridyl moiety exposed on the kinase surface. Mutations of single residues that are exposed on the protein surface have been shown to promote the formation of higher-order protein complexes; the haemoglobin Glu to Val mutation, for example, induces polymerization in sickle cell anaemia[32]. Accordingly, single-residue mutations that are designed to increase surface hydrophobicity give rise to ordered protein assemblies[33]. Bound compounds—for example, enzyme inhibitors—can in principle mimic such amino acid changes and thereby have strong effects on the protein interaction landscape, suggesting that compound-induced protein–protein interactions may be more common than previously recognized. Together, our results suggest that the modification of surface-exposed regions in target-binding small molecules is a rational strategy that could be used to develop molecular glue degraders for a given protein target.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2374-x.

1. Stanton, B. Z., Chory, E. J. & Crabtree, G. R. Chemically induced proximity in biology and medicine. *Science* **359**, eaao5902 (2018).
2. Chopra, R., Sadok, A. & Collins, I. A critical evaluation of the approaches to targeted protein degradation for drug discovery. *Drug Discov. Today. Technol.* **31**, 5–13 (2019).
3. Yu, C. et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat. Biotechnol.* **34**, 419–423 (2016).
4. Corsello, S. M. et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
5. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
6. Bettayeb, K. et al. CR8, a potent and selective, roscovitine-derived inhibitor of cyclin-dependent kinases. *Oncogene* **27**, 5797–5807 (2008).
7. Fink, E. C. & Ebert, B. L. The novel mechanism of lenalidomide activity. *Blood* **126**, 2366–2369 (2015).
8. Lu, G. et al. The myeloma drug lenalidomide promotes the cereblon-dependent destruction of Ikaros proteins. *Science* **343**, 305–309 (2014).
9. Krönke, J. et al. Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301–305 (2014).
10. Matyskiela, M. E. et al. A novel cereblon modulator recruits GSPT1 to the $CRL4^{CRBN}$ ubiquitin ligase. *Nature* **535**, 252–257 (2016).
11. Petzold, G., Fischer, E. S. & Thomä, N. H. Structural basis of lenalidomide-induced CK1α degradation by the $CRL4^{CRBN}$ ubiquitin ligase. *Nature* **532**, 127–130 (2016).
12. Fischer, E. S. et al. Structure of the DDB1-CRBN E3 ubiquitin ligase in complex with thalidomide. *Nature* **512**, 49–53 (2014).
13. Han, T. et al. Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. *Science* **356**, eaal3755 (2017).
14. Uehara, T. et al. Selective degradation of splicing factor CAPERα by anticancer sulfonamides. *Nat. Chem. Biol.* **13**, 675–680 (2017).
15. Faust, T. B. et al. Structural complementarity facilitates E7820-mediated degradation of RBM39 by DCAF15. *Nat. Chem. Biol.* **16**, 7–14 (2020).
16. Bondeson, D. P. et al. Catalytic in vivo protein knockdown by small-molecule PROTACs. *Nat. Chem. Biol.* **11**, 611–617 (2015).
17. Angers, S. et al. Molecular architecture and assembly of the DDB1-CUL4A ubiquitin ligase machinery. *Nature* **443**, 590–593 (2006).
18. Lee, J. & Zhou, P. DCAFs, the missing link of the CUL4-DDB1 ubiquitin ligase. *Mol. Cell* **26**, 775–780 (2007).
19. Bösken, C. A. et al. The structure and substrate specificity of human Cdk12/Cyclin K. *Nat. Commun.* **5**, 3505 (2014).
20. Cheng, S.-W. G. et al. Interaction of cyclin-dependent kinase 12/CrkRS with cyclin K1 is required for the phosphorylation of the C-terminal domain of RNA polymerase II. *Mol. Cell. Biol.* **32**, 4691–4704 (2012).
21. Dixon-Clarke, S. E., Elkins, J. M., Cheng, S. W. G., Morin, G. B. & Bullock, A. N. Structures of the CDK12/CycK complex with AMP-PNP reveal a flexible C-terminal kinase extension important for ATP binding. *Sci. Rep.* **5**, 17122 (2015).
22. Oumata, N. et al. Roscovitine-derived, dual-specificity inhibitors of cyclin-dependent kinases and casein kinases 1. *J. Med. Chem.* **51**, 5229–5242 (2008).
23. Meijer, L. et al. Biochemical and cellular effects of roscovitine, a potent and selective inhibitor of the cyclin-dependent kinases cdc2, cdk2 and cdk5. *Eur. J. Biochem.* **243**, 527–536 (1997).
24. Zhang, T. et al. Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* **12**, 876–884 (2016).
25. Sedlacek, H. et al. Flavopiridol (L86 8275; NSC 649890), a new kinase inhibitor for tumor therapy. *Int. J. Oncol.* **9**, 1143–1168 (1996).
26. Sievers, Q. L. et al. Defining the human C2H2 zinc finger degrome targeted by thalidomide analogs through CRBN. *Science* **362**, eaat0572 (2018).
27. Jones, L. H. Small-molecule kinase downregulators. *Cell Chem. Biol.* **25**, 30–35 (2018).
28. Schreiber, S. L. A chemical biology view of bioactive small molecules and a binder-based approach to connect biology to precision medicines. *Isr. J. Chem.* **59**, 52–59 (2019).
29. Ito, T. et al. Identification of a primary target of thalidomide teratogenicity. *Science* **327**, 1345–1350 (2010).
30. Fischer, E. S. et al. The molecular basis of $CRL4^{DDB2/CSA}$ ubiquitin ligase architecture, targeting, and activation. *Cell* **147**, 1024–1039 (2011).
31. Johnson, S. F. et al. CDK12 inhibition reverses de novo and acquired PARP inhibitor resistance in BRCA wild-type and mutated models of triple-negative breast cancer. *Cell Rep.* **17**, 2367–2381 (2016).
32. Dykes, G. W., Crepeau, R. H. & Edelstein, S. J. Three-dimensional reconstruction of the 14-filament fibers of hemoglobin S. *J. Mol. Biol.* **130**, 451–472 (1979).
33. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548**, 244–247 (2017).

Methods

Data reporting
No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Mammalian cell culture
The human HEK293T cell lines were provided by the Genetic Perturbation Platform, Broad Institute and the K562-Cas9, THP1-Cas9 and P31FUJ-Cas9 cell lines were provided by Z. Tothova (Broad Institute). MOLT-4 cells were purchased from ATCC and HEK293T-Cas9 cells[26] and MM1S-Cas9 cells[34] were previously published. Sf9 cells were purchased from Thermo Fisher Scientific and Hi5 cells were purchased from Expression Systems. HEK293T, K562-Cas9, THP1-Cas9, P31FUJ-Cas9, HEK293T-Cas9, MM1S-Cas9 and MOLT-4 cell lines were mycoplasma-negative and authenticated by STR profiling. Sf9 and Hi5 cells were authenticated by the vendor. HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM (Gibco) and all other cell lines in RPMI (Gibco), with 10% fetal bovine serum (FBS) (Invitrogen), glutamine (Invitrogen) and penicillin–streptomycin (Invitrogen) at 37 °C and 5% CO_2.

Compounds
(*R*)-CR8 (3605) was obtained from Tocris, (*S*)-CR8 (ALX-270-509-M005) and flavopiridol (ALX-430-161-M005) from Enzo Life Sciences, roscovitine (HY-30237), THZ531 (HY-103618) and LDC00067 (HY-15878) from MedChem Express and DRF053 (D6946) from Sigma.

Antibodies
The following antibodies were used in this study: anti-cyclin K (Bethyl Laboratories, A301-939A for full length cyclin K), anti-cyclin K (Abcam, ab251652, for cyclin K(1–267), anti-β-actin (Cell Signaling, 3700), anti-CRBN (Sigma Prestige, HPA045910), anti-mouse 800CW (LI-COR Biosciences, 926-32211), anti-rabbit 680LT (LI-COR Biosciences, 925-68021) and anti-rabbit IgG antibodies (Abcam, ab6721).

Reporter vectors
The following reporters were used in this study: Artichoke (SFFV.BsmBICloneSite-17aaRigidLinker-eGFP.IRES.mCherry.cppt.EF1α.PuroR, Addgene 73320) for genome-wide screen and validation experiments; Cilantro 2 (PGK.BsmBICloneSite-10aaFlexibleLinker-eGFP.IRES.mCherry. cppt.EF1α.PuroR, Addgene 74450) for degradation kinetics; sgBFP (sgRNA.SFFV.tBFP) for validation of drug–E3 ligase pairs; sgRFP657 (sgRNA.EFS.RFP657) for validation of drug–E3 ligase pairs; and sgPuro (pXPR003, Addgene 52963) for drug-sensitivity assays.

Oligonucleotides
List of all oligonucleotides used in this study can be found in Supplementary Table 1.

Bioinformatic screen
We computed Pearson correlations of the toxicity of PRISM repurposing compounds in 8 doses and 578 cell lines[4] with gene expression and copy-number variation of all detectable protein-coding genes of matched cell lines from The Cancer Cell Line Encyclopedia (CCLE)[5]. A *z* score was computed for each pair of compounds, dose toxicity and genomic feature (gene expression or copy-number variation) across all cell lines. For each compound–genomic feature pair, the most extreme correlations are ranked from negative to positive. To focus on novel relationships between compounds and genes, we restricted genes to a curated list of 499 E3 ligase components and compounds that are not an 'EGFR inhibitor', 'RAF inhibitor' or 'MDM inhibitor' on the basis of PRISM repurposing annotation[5]. Hit compounds were selected if either the *z* score was less than −6 or the compound was ranked in the top 15 with a *z* score less than −4. The resulting list of 158 E3 gene–compound pairs was further curated and shortened manually to 96 E3 gene–compound pairs, which included 95 unique E3 ligases and 85 unique compounds.

Cloning and lentiviral packaging of sgRNAs targeting 95 E3 ligases
sgRNAs targeting E3 ligases were selected from the human Brunello CRISPR library[35]. A total of 170 pairs of oligonucleotides (IDT) targeting 95 E3 ligases were annealed and cloned into the sgRNA.SFFV.tBFP (guide ID A) or sgRNA.EFS.RFP657 (guide ID B) fluorescent vectors in a 96-well format using previously published protocols[36]. In brief, vectors were linearized with BsmBI (New England Biolabs) and gel-purified with the Spin Miniprep Kit (Qiagen). Annealed oligonucleotides were phosphorylated with T4 polynucleotide kinase (New England Biolabs) and ligated into the linearized and purified vector backbones with T4 DNA ligase (New England Biolabs). Constructs were transformed into XL10-Gold ultracompetent *Escherichia coli* (Stratagene/Agilent Technologies), plasmids were purified using the Miniprep Kit (Qiagen), and the guide sequence was confirmed by Sanger sequencing. For validation of the primary screen, virus was produced in a 96-well format. In brief, 11,000 HEK293T cells were seeded per well in 100 µl DMEM medium supplemented with 10% FBS and penicillin–streptomycin–glutamine. The next day a packaging mix was prepared in a 96-well plate consisting of 500 ng psPAX2, 50 ng pVSV-G and 17 ng sgRNA backbone in 5 µl OptiMem (Invitrogen) and incubated for 5 min at room temperature. This mix was combined with 0.1 µl TransIT-LT1 (Mirus) in 5 µl OptiMem, incubated for 30 min at room temperature and then applied to cells. Two days after transfection, dead cells were removed by centrifugation and lentivirus-containing medium was collected and stored at −80 °C before use.

Validation of drug–E3 ligase pairs from the bioinformatic screen
K562-Cas9, OVK16-Cas9, A564-Cas9, ES2-Cas9 and MOLM13-Cas9 cell lines were individually transduced with 192 sgRNAs targeting 95 E3 ligases in a 96-well plate format. Exactly 3,000 cells per well were plated in 100 µl RPMI supplemented with 10% fetal calf serum (FCS) and penicillin–streptomycin–glutamine and 30 µl per well of virus supernatant was added. The medium was changed 24 hours after infection. After three days, the percentage of sgRNA-transduced cells was determined by flow cytometry. If more than 60% of cells were transduced, untransduced cells were added to bring the level below 60%. Eight days after infection, the cell density was measured and adjusted to 1.5×10^5 cells per ml with RPMI. For treatment, 50 µl of sgRNA-transduced cells were seeded into each well of a 384-well plate with preplated DMSO or cognate drug in three concentrations (0.1 µM, 1 µM or 10 µM) with the Agilent BRAVO Automated Liquid Handling Platform. Plates were sealed with White Rayon adhesive sealing tape (Thermo Fisher Scientific) and grown for three days. Adherent cell lines were trypsinized and resuspended in 50 µl RPMI with Matrix WellMate (Thermo Fisher Scientific). Suspension cells were directly subjected to analysis. Around 10 µl of cell suspension was subjected to flow analysis with a FACSCanto equipped with a high throughput sampler (BD Biosciences). The percentage of sgRNA-transduced cells in the drug-treatment wells was normalized to the DMSO control. Wells with fluorescent drug and samples with fewer than 120 viable cells or less than 6% fluorescent cells were removed from the analysis. All E3–drug pairs were ranked on the basis of the number of experimental conditions (cell line and drug dose) with more than 50% of sgRNA-transduced cells in drug-treatment wells in comparison with the corresponding DMSO control wells.

Validation of DDB1-resistance phenotype
For validation experiments, virus was produced in a six-well plate format, as described above with the following adjustments: 2.5×10^5 HEK293T cells per well in 2 ml DMEM medium, 3 µl per well TransIT-LT1,

15 µl per well OPTI-MEM, 500 ng per well of the desired plasmid, 500 ng per well psPAX2 and 50 ng per well pVSV-G in 32.5 µl per well OPTI-MEM. After collecting the virus, 10×10^3 HEK293T-Cas9 cells in 100 µl DMEM medium were transduced with 10 µl of virus supernatant. The transduced HEK293T-Cas9 cells were then mixed with untransduced control cells in a 1:9 ratio. Nine days after sgRNA transduction, cells were treated for three days with DMSO or 1 µM CR8 and analysed by flow cytometry to determine the percentage of BFP^+ cells. sgRNAs targeting *DDB1* provide partial depletion of DDB1 (50% *DDB1* alleles modified, reducing DDB1 levels by roughly 50%), which suggests selection towards heterozygous or hypomorphic clones.

Whole-proteome quantification using tandem mass tag mass spectrometry

Around 10×10^6 MOLT-4 cells were treated with DMSO (triplicate) or 1 µM CR8 (single replicate) for 1 h or 5 h and later were collected by centrifugation. Samples were processed, measured and analysed as described before[37]. Data are available in the PRIDE repository (PXD016187 and PXD016188).

Quantitative PCR

HEK293T-Cas9 cells were treated with DMSO or 1 µM CR8 for 2 h, collected by centrifugation, washed with phosphate-buffered saline (PBS) and snap-frozen at −80 °C. mRNA was isolated using a QIAGEN RNA kit (Qiagen, 74106). For cDNA synthesis, total RNA was reverse-transcribed using a High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific) before quantitative PCR (qPCR) analysis with TaqMan Fast Advanced Master Mix (Thermo Fisher Scientific, 4444557) for *CCNK* (TaqMan, Hs00171095_m1, Life Technologies) and *GAPDH* (TaqMan, Hs02758991_g1). Reactions were run and analysed on a CFX96 Real Time system (Bio-Rad).

Immunoblots for whole-protein lysate

Cells were washed with PBS and lysed (150 mM NaCl, 50 mM Tris (pH 7.5), 1% NP-40, 1% glycerol, 1× Halt Cocktail protease and phosphatase inhibitors) for 20 min on ice. The insoluble fraction was removed by centrifugation, the protein concentration was quantified using a BCA protein assay kit (Pierce), and an equal amount of lysate was run on SDS–PAGE 4–12% Bis–Tris Protein Gels (NuPAGE, Thermo Fisher Scientific) and then transferred to nitrocellulose membrane with a Trans-Blot Turbo System (Bio-Rad). Membranes were blocked in Odyssey Blocking Buffer/PBS (LI-COR Biosciences) and incubated with primary antibodies overnight at 4 °C. The membranes were then washed in Tris-buffered saline with Tween 20 (TBS-T), incubated for 1 h with secondary IRDye-conjugated antibodies (LI-COR Biosciences) and washed three times in TBS-T for 5 min before near-infrared western blot detection on an Odyssey Imaging System (LI-COR Biosciences).

Cyclin K reporter stability analysis

HEK293T-Cas9 cells expressing the cyclin K_{eGFP} degradation reporter were transduced with experimental sgRNAs. Nine days after infection, the cells were dosed for 2 h with DMSO or 1 µM CR8 and the fluorescent signal was quantified by flow cytometry (CytoFLEX, Beckman or LSR-Fortessa flow cytometer, BD Biosciences). Using FlowJo (flow cytometry analysis software, BD Biosciences), the geometric mean of the eGFP and mCherry fluorescent signal for round and mCherry-positive cells was calculated. The ratio of eGFP to mCherry was normalized to the average of three DMSO-treated controls.

Genome-wide CRISPR screen for CR8 resistance

Five per cent (v/v) of the human genome-wide CRISPR-KO Brunello library with 0.4 µl polybrene ml^{-1} (stock of 8 mg ml^{-1}) was added to 1.5×10^8 HEK293T-Cas9 cells in 75 ml medium and transduced (2,400 rpm, 2 h, 37 °C). Twenty-four hours after infection, sgRNA-transduced cells were selected with 2 µg ml^{-1} puromycin for two days. On the ninth day after infection, cells were treated with either DMSO ($n = 1$) or 1 µM CR8 ($n = 1$) and cultured for an additional three days. Resistant live cells were selected by gently washing away detached dead cells from the medium. Cell pellets were resuspended in multiple direct lysis buffer reactions (1 mM $CaCl_2$, 3 mM $MgCl_2$, 1 mM EDTA, 1% Triton X-100, Tris pH 7.5, with freshly supplemented 0.2 mg ml^{-1} proteinase K) with 1×10^6 cells per 100-µl reaction. The sgRNA sequence was amplified in a first PCR reaction with eight staggered forward primers. Direct lysed cells (20 µl) were mixed with 0.04U Titanium Taq (Takara Bio 639210), 0.5× Titanium Taq buffer, 800 µM dNTP mix, 200 nM SBS3-Stagger-pXPR003 forward primer and 200 nM SBS12-pXPR003 reverse primer in a 50-µl reaction (cycles: 5 min at 94 °C, 15 × (30 s at 94 °C, 15 s at 58 °C, 30 s at 72 °C), 2 min at 72 °C). Exactly 2 µl of the first PCR reaction was used as the template for 15 cycles of the second PCR, in which Illumina adapters and barcodes were added (0.04U Titanium Taq, 1× Titanium Taq buffer, 800 µM dNTP mix, 200 nM P5-SBS3 forward primer, 200 nM P7-barcode-SBS12 reverse primer). An equal amount of all samples was pooled and subjected to preparative agarose electrophoresis followed by gel purification (Qiagen). Eluted DNA was further purified by NaOAc and isopropanol precipitation. Amplified sgRNAs were quantified using the Illumina NextSeq platform (Genomics Platform, Broad Institute). Read counts for all guides targeting the same gene were used to generate *P* values. Hits enriched in the resistant population with a false discovery rate (FDR) < 0.05 and enriched more than fivefold are labelled on the plot[38] (Fig. 1f).

Bison CRISPR screen for CR8 resistance

The Bison CRISPR library targets 713 E1, E2 and E3 ubiquitin ligases, deubiquitinases and control genes and contains 2,852 guide RNAs. It was cloned into pXPR003 as previously described[35] by the Broad Institute Genetic Perturbation Platform (GPP). The virus for the library was produced in a T-175 flask format, as described above with the following adjustments: 1.8×10^7 HEK293T cells in 25 ml complete DMEM medium, 244 µl TransIT-LT1, 5 ml OPTI-MEM, 32 µg library, 40 µg psPAX2 and 4 µg pVSV-G in 1 ml OPTI-MEM. Ten per cent (v/v) of the Bison CRISPR library was added to 6×10^6 HEK293T-Cas9 cells in triplicates and transduced. Samples ($n = 3$) were processed as described above for the genome-wide resistance screen.

Genome-wide CRISPR screen for cyclin K reporter stability

A single clone of cyclin K_{eGFP} HEK293T-Cas9 was transduced with the genome-wide Brunello library as described above with the following modification: 4.5×10^8 cyclin K_{eGFP} HEK293T-Cas9 cells in 225 ml medium. Nine days later, cells were treated with CR8 ($n = 3$) or DMSO ($n = 3$) for at least 2 h and the cyclin K stable population was separated using fluorescence-activated cell sorting (FACS). Four populations were collected (top 5%, top 5–15%, lowest 5–15% and lowest 5%) on the basis of the cyclin K_{eGFP} to mCherry mean fluorescent intensity (MFI) ratio on an MA900 Cell Sorter (Sony). Sorted cells were collected by centrifugation and subjected to direct lysis as described above. The screen was analysed as described below by comparing stable populations (top 5% eGFP/mCherry expression) to unstable populations (lowest 15% eGFP/mCherry expression). Hits enriched in the cyclin K stable population with FDR < 0.05 are labelled on the plot (Fig. 1g).

Data analysis of the pooled CRISPR screens

The data analysis pipeline comprised the following steps. (1) Each sample was normalized to the total number of reads. (2) For each guide, the ratio of reads in the stable versus the unstable sorted gate was calculated, and sgRNAs were ranked. (3) The ranks for each guide were summed for all replicates. (3) The gene rank was determined as the median rank of the four guides targeting it. (4) *P* values were calculated by simulating a distribution with guide RNAs that had randomly assigned ranks over 100 iterations. R scripts can be found in the Supplementary Information.

Screen with arrayed DCAF library

An arrayed DCAF library (targeting DCAF substrate receptors, DCAF-like and control genes) was constructed as described above with the appropriate oligonucleotides (Supplementary Table 1). K562-Cas9, P31FUJ-Cas9, THP1-Cas9 and MM1S-Cas9 cells were individually transduced and treated with DMSO or 1 μM CR8 (K562-Cas9, P31FUJ-Cas9, THP1-Cas9) or 0.1 μM CR8 (MM1S-Cas9). The analysis was performed as described above for the validation of the DDB1-resistance phenotype.

Protein purification

Human wild-type and mutant versions of DDB1 (Uniprot entry Q16531), CDK12 (Q9NYV4, K965R) and cyclin K (O75909) were subcloned into pAC-derived vectors[39] and recombinant proteins were expressed as N-terminal His_6, His_6–Spy, StrepII or StrepII–Avi fusions in *Trichoplusia ni* High Five insect cells using the baculovirus expression system (Invitrogen)[40].

Wild-type or mutant forms of full-length or BPB domain deletion (ΔBPB: amino acids 396–705 deleted) constructs of His_6–DDB1 and StrepII–Avi–DDB1 were purified as previously described for DDB1–DCAF complexes[12]. High Five insect cells co-expressing truncated versions of wild-type or mutant His_6–CDK12 (amino acids 713–1052 or 713–1032) and His_6- or His_6–Spy-tagged cyclin K (amino acids 1–267) were lysed by sonication in 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10% (v/v) glycerol, 10 mM $MgCl_2$, 10 mM imidazole, 0.25 mM tris(2-carboxyethyl)phosphine (TCEP), 0.1% (v/v) Triton X-100, 1 mM phenylmethylsulfonylfluoride (PMSF) and 1× protease inhibitor cocktail (Sigma). After ultracentrifugation, the soluble fraction was passed over HIS-Select Ni^{2+} affinity resin (Sigma), washed with 50 mM Tris-HCl (pH 8.0), 1 M NaCl, 10% (v/v) glycerol, 0.25 mM TCEP and 10 mM imidazole and eluted in 50 mM Tris-HCl (pH 8.0), 200 mM NaCl, 10% (v/v) glycerol, 0.25 mM TCEP and 250 mM imidazole. When necessary, affinity tags were removed by overnight tobacco etch virus (TEV) protease treatment. In cases of HIS-Select Ni^{2+} affinity-purified CDK12–cyclin K that was not subjected to TEV cleavage, the pH of the eluate was adjusted to 6.8 before ion-exchange chromatography. StrepII-tagged versions of CDK12–cyclin K were affinity-purified using Strep-Tactin Sepharose (IBA), omitting imidazole in lysis, wash and elution buffers, supplementing the elution buffer with 2.5 mM desthiobiotin (IBA GmbH) and using 50 mM Tris-HCl (pH 6.8) throughout.

For ion-exchange chromatography, affinity-purified proteins were diluted in a 1:1 ratio with buffer A (50 mM Tris-HCl (pH 6.8), 10 mM NaCl, 2.5% (v/v) glycerol and 0.25 mM TCEP) and passed over an 8-ml Poros 50HQ column. The flow-through was again diluted in a 1:1 ratio with buffer A and passed over an 8-ml Poros 50HS column. Bound proteins were eluted by a linear salt gradient mixing buffer A and buffer B (50 mM Tris-HCl (pH 6.8), 1 M NaCl, 2.5% (v/v) glycerol and 0.25 mM TCEP) over 15 column volumes to a final ratio of 80% buffer B. Poros 50HS peak fractions containing the CDK12–cyclin K complex were concentrated and subjected to size-exclusion chromatography in 50 mM HEPES (pH 7.4), 200 mM NaCl, 2.5% (v/v) glycerol and 0.25 mM TCEP. The concentrated proteins were flash-frozen in liquid nitrogen and stored at −80 °C.

Co-immunoprecipitation assay

The purified $CDK12_{His6}$–cyclin $K_{StrepII}$ complex was mixed with equimolar concentrations of full-length His_6-tagged DDB1 or TEV-cleaved DDB1(ΔBPB) (5 μM) in the presence of 5 μM (*R*)-CR8 or DMSO in immunoprecipitation (IP) buffer (50 mM HEPES (pH 7.4), 200 mM NaCl, 0.25 mM TCEP and 0.05% (v/v) Tween-20) containing 1 mg ml^{-1} bovine serum albumin. The solution was added to Strep-Tactin MacroPrep beads (IBA GmbH) pre-equilibrated in IP buffer and incubated for 1 h at 4 °C on an end-over-end shaker. The beads were extensively washed with IP buffer, and the bound protein was eluted with IP buffer containing 2.5 mM desthiobiotin for 1 h at 4 °C on an end-over-end shaker. Eluted proteins were separated by SDS–PAGE and stained with Coomassie blue.

Crystallization and data collection

The protein solution for crystallization contained 70 μM TEV-cleaved DDB1(ΔBPB), 80 μM (*R*)-CR8 and 80 μM TEV-cleaved CDK12–cyclin K in 50 mM HEPES (pH 7.4), 200 mM NaCl and 0.25 mM TCEP. Crystals were grown by vapour diffusion in drops containing 200 nl DDB1(ΔBPB)–(*R*)-CR8–CDK12(713–1052)–cyclin K(1–267) complex solution mixed with 200 nl of reservoir solution containing 0.9 M ammonium citrate tribasic (pH 7.0) in two-well-format sitting drop crystallization plates (Swissci). Plates were incubated at 19 °C and crystals appeared 5–13 days after set-up. Crystals were flash-cooled in liquid nitrogen in reservoir solution supplemented with 25% (v/v) glycerol as a cryoprotectant before data collection. Diffraction data were collected at the Swiss Light Source (SLS; beamline PXI) with an Eiger 16M detector (Dectris) at a wavelength of 1 Å and a crystal cooled to 100 K. Data were processed with DIALS, scaled with AIMLESS supported by other programs of the CCP4 suite[41] and converted to structure factor amplitudes with STARANISO[42], applying a locally weighted $CC_{1/2} = 0.3$ resolution cut-off.

Structure determination and model building

The DDB1(ΔBPB)–(*R*)–CR8–CDK12(713–1052)–cyclin K(1–267) complex formed crystals belonging to space group $P3_121$, with three complexes in the crystallographic asymmetric unit. Their structure was determined using molecular replacement in Phaser[43] with a search model derived from PDB entry 6H0F for DDB1(ΔBPB) and PDB entry 4NST for CDK12–cyclin K. The initial model was improved by iterative cycles of building with Coot[44], and refinement using phenix.refine[45] or autoBUSTER[46], with ligand restraints generated using eLBOW through phenix.ready_set[47]. The final model was produced by refinement with autoBUSTER. Analysis with MolProbity[48] indicates that 93.9% of the residues in the final model are in favourable regions of the Ramachandran plot, with 0.6% outliers. Data processing and refinement statistics are provided in Extended Data Table 1. Interface analysis was performed using PISA[49].

Biotinylation of DDB1

Purified full-length StrepII–Avi–DDB1 was biotinylated in vitro at a concentration of 8 μM by incubation with final concentrations of 2.5 μM BirA enzyme and 0.2 mM D-biotin in 50 mM HEPES (pH 7.4), 200 mM NaCl, 10 mM $MgCl_2$, 0.25 mM TCEP and 20 mM ATP. The reaction was incubated for 1 h at room temperature and stored at 4 °C for 14–16 h. Biotinylated DDB1 ($DDB1_{biotin}$) was purified by gel-filtration chromatography and stored at −80 °C (around 20 μM).

TR-FRET

Increasing concentrations of Alexa488–SpyCatcher-labelled[26] His_6–Spy–cyclin K in complex with His_6–CDK12 ($CDK12–cyclinK_{Alexa488}$) were added to a mixture of $DDB1_{biotin}$ at 50 nM, terbium-coupled streptavidin at 4 nM (Invitrogen) and compounds at 10 μM (final concentrations) in 384-well microplates (Greiner, 784076) in a buffer containing 50 mM Tris (pH 7.5), 150 mM NaCl, 0.1% pluronic acid and 0.5% DMSO (see also figure legends). Before TR-FRET measurements, reactions were incubated for 15 min at room temperature. After excitation of terbium fluorescence at 337 nm, emissions at 490 nm (terbium) and 520 nm (Alexa488) were measured with a 70-μs delay to reduce background fluorescence and the reactions were followed by recording 60 data points of each well over 1 h using a PHERAstar FS microplate reader (BMG Labtech). The TR-FRET signal of each data point was extracted by calculating the 520:490 nm ratio. Data were analysed with Prism 7 (GraphPad) assuming equimolar binding of $DDB1_{biotin}$ to $CDK12–cyclinK_{Alexa488}$ using the equations described previously[8].

Counter-titrations with unlabelled proteins were carried out by mixing 500 μM $CDK12–cyclinK_{Alexa488}$ with 50 nM $DDB1_{biotin}$ in the presence of 4 nM terbium-coupled streptavidin and 1 μM compound for

titrations with unlabelled DDB1 or 12.5 μM compound for titrations with unlabelled CDK12. After incubation for 15 min at room temperature, increasing amounts of unlabelled CDK12–cyclin K or DDB1 (0–10 μM) were added to the preassembled CDK12–cyclin$K_{Alexa488}$–$DDB1_{biotin}$ complexes in a 1:1 volume ratio and incubated for 15 min at room temperature. TR-FRET data were acquired as described above. The 520:490 nm ratios were plotted to calculate the half maximal inhibitory concentrations (IC_{50}) assuming a single binding site using Prism 7 (GraphPad). IC_{50} values were converted to the respective inhibition constant (K_i) values as described previously[50]. Three technical replicates were carried out per experiment.

CUL4–RBX1–DDB1 reconstitution and in vitro CUL4 neddylation

In vitro CRL4 reconstitution and CUL4 neddylation were performed as described[11]. $CUL4A_{His6}$–$RBX1_{His6}$ at 3.5 μM was incubated with His_6-tagged DDB1 at 3 μM in a reaction mixture containing 3.8 μM NEDD8, 50 nM NAE1–UBA3 (E1), 30 nM UBC12 (E2), 1 mM ATP, 50 mM Tris (pH 7.5), 100 mM NaCl, 2.5 mM $MgCl_2$, 0.5 mM DTT and 5% (v/v) glycerol for 1.5 h at room temperature. Neddylated and gel-filtration-purified CUL4–RBX1–DDB1 ($CUL4_{NEDD8}$–RBX1–DDB1) was concentrated to 7.6 μM, flash-frozen and stored at −80 °C.

In-vitro ubiquitination assays

In vitro ubiquitination was performed by mixing $CUL4_{NEDD8}$–RBX1–DDB1 at 70 nM with a reaction mixture containing compounds at 2 μM, CDK12–cyclin K at 500 nM, E1 (UBA1, Boston Biochem) at 50 nM, E2 (UBCH5a, Boston Biochem) at 1 μM and ubiquitin at 20 μM. Reactions were carried out in 50 mM Tris (pH 7.5), 150 mM NaCl, 5 mM $MgCl_2$, 0.2 mM $CaCl_2$, 1 mM ATP, 0.1% Triton X-100 and 0.1 mg ml^{-1} bovine serum albumin (BSA), incubated for 0–30 min at 30 °C and analysed by western blot using anti-cyclin K and anti-rabbit IgG antibodies. Blots were scanned on an Amersham 600 CCD-based imaging system (GE Life Sciences).

ITC

ITC experiments were performed at 25 °C on a VP-ITC isothermal titration calorimeter (Microcal). Purified and TEV-cleaved CDK12–cyclin K and DDB1(ΔBPB) were exhaustively dialysed in 50 mM HEPES (pH 7.4), 150 mM NaCl, 0.25 mM TCEP and 0.5% DMSO and loaded into the sample cell at a final concentration of 10–50 μM. Kinase inhibitors (CR8 or roscovitine) were diluted from a 100 mM DMSO stock solution to 100–500 μM in buffer containing 50 mM HEPES (pH 7.4), 150 mM NaCl and 0.25 mM TCEP. The final DMSO concentration was 0.5%. Titrations with 100–500 μM compound were performed typically through about 30 injections of 6–10 μl at 210-s intervals from a 300-μl syringe rotating at 300 rpm. An initial injection of the ligand (4 μl) was made and discarded during data analysis. For probing DDB1–CDK12–cyclin K complex formation, DDB1(ΔBPB) (100 μM, in the syringe) was titrated into the cell containing CDK12–cyclin K (10 μM) or CDK12–cyclin K (10 μM) pre-incubated with CR8 (30 μM). The heat change accompanying the titration was recorded as differential power by the instrument and determined by integration of the peak obtained. Titrations of ligand to buffer only and buffer into protein were performed to allow baseline corrections. The heat change was fitted using nonlinear least-squares minimization to obtain the dissociation constants (K_d), the enthalpy of binding (ΔH) and the stoichiometry (N). Between one and three replicates were performed per titration.

Bioluminescence resonance energy transfer analyses

Bioluminescence resonance energy transfer (BRET) experiments were performed using a NanoBRET PPI starter kit (Promega N1821) according to the manufacturer's instructions and as previously described[51].

Drug-sensitivity assays

HEK293T-Cas9 cells were resuspended at 0.15×10^6 per ml and plated on a 384-well plate with 50 μl per well and MLN4924, MLN7243 or MG132 with or without CR8 serially diluted with D300e Digital Dispenser (Tecan).

HEK293T-Cas9 cells (0.625×10^6 cells per well of a 6-well plate) were seeded the day before transfection. The following day, 2.5 ug of pRSF91-GFP or pRSF91-CRBN[9] plasmid DNA was mixed with 250 μl Opti-Mem and 7.5 μl TransIT-LT1 (Mirus Bio) according to the manufacturer's protocol. Two days (48 h) after transfection, cells were resuspended at 0.15×10^6 cells per ml and plated on a 384-well plate with 50 μl per well.

HEK293T-Cas9 cells were transduced with sgRNAs targeting either *DDB1* or luciferase in a pXPR003 backbone (GPP) (Supplementary Table 1). After nine days of puromycin selection, cells were replated into a 96-well format with 2×10^4 cells per well and CR8 and roscovitine were serially diluted with D300e Digital Dispenser (Tecan).

After three days of drug exposure, cell viability was assessed using the CellTiter-Glo luminescent assay (Promega, G7572) on an EnVision Multilabel Plate Reader (Perkin Elmer) or CLARIOstar Plus, MARS 3.4 (BMG LabTech). Cell viabilities were calculated relative to DMSO controls. The half-maximum effective concentration (EC_{50}) values were derived from standard four-parameter log-logistic curves fitted with the 'dr4pl' R package.

Cyclin K reporter stability analysis with CRBN overexpression

HEK293T-Cas9 cells expressing the cyclin K_{eGFP} degradation reporter were transiently transfected with pLX307-Luc or pLX307-CRBN (for flow experiment) as described above and 48 h after infection were treated with CR8 for 2 h and analysed by flow cytometry.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Structural data have been deposited in the PDB under the accession code 6TD3. Proteome quantification data are available in the PRIDE repository (PXD016187 and PXD016188). Additional ITC data are provided in Supplementary Fig. 1. Uncropped gel and western blot source data are shown in Supplementary Fig. 2 and the flow cytometry gating strategy is shown in Supplementary Fig. 3.

Code availability

The code necessary to reproduce the statistical analysis is included in the Supplementary Information.

34. Sievers, Q. L., Gasser, J. A., Cowley, G. S., Fischer, E. S. & Ebert, B. L. Genome-wide screen identifies cullin-RING ligase machinery required for lenalidomide-dependent $CRL4^{CRBN}$ activity. *Blood* **132**, 1293–1303 (2018).
35. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
36. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
37. Donovan, K. A. et al. Thalidomide promotes degradation of SALL4, a transcription factor implicated in Duane Radial Ray syndrome. *eLife* **7**, e38430 (2018).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
39. Abdulrahman, W. et al. A set of baculovirus transfer vectors for screening of affinity tags and parallel expression strategies. *Anal. Biochem.* **385**, 383–385 (2009).
40. Li, T., Pavletich, N. P., Schulman, B. A. & Zheng, N. High-level expression and purification of recombinant SCF ubiquitin ligases. *Methods Enzymol.* **398**, 125–142 (2005).
41. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D.* **67**, 235–242 (2011).
42. Tickle, I. J. et al. *STARANISO* (Global Phasing, 2018).
43. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
44. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D* **66**, 486–501 (2010).
45. Afonine, P. V. et al. Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr. D* **68**, 352–367 (2012).
46. Bricogne, G. et al. *BUSTER* (Global Phasing, 2017).

47. Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D* **65**, 1074–1080 (2009).
48. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
49. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
50. Cer, R. Z., Mudunuri, U., Stephens, R. & Lebeda, F. J. IC_{50}-to-K_i: a web-based tool for converting IC_{50} to K_i values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Res.* **37**, W441–W445 (2009).
51. Sperling, A. S. et al. Patterns of substrate affinity, competition, and degradation kinetics underlie biological activity of thalidomide analogs. *Blood* **134**, 160–170 (2019).

Acknowledgements We thank S. Cavadini and A. Schenk for support during electron microscopy data collection and processing; M. Kolesnikov for help with ITC experiments; the Broad Institute PRISM (particularly M. Kocak); the Compound Management, Cancer Data Science, Walk-up Sequencing, Genetic Perturbation Platform and Flow Facility teams (particularly P. Rogers), K. DeRuff for the introduction to the Agilent BRAVO Automated Liquid Handling Platform, and J. Kennedy for providing sgRNA.SFFV.tBFP and sgRNA.EFS.RFP657 backbones. We acknowledge the Paul Scherrer Institute for provision of synchrotron radiation beam time at beamline PXI of the SLS and thank T. Tomizaki for assistance. We are grateful to all members of the Ebert group, particularly R. Belizaire, S. Koochaki, P. Miller and C. Zou; as well as R. Nowak and P. Tsvetkov, for discussions. This work was supported by the NIH (grants R01HL082945, P01CA108631 and P50CA206963), the Howard Hughes Medical Institute, the Edward P. Evans Foundation and the Leukaemia and Lymphoma Society (to B.L.E.); funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 666068, the Gebert Rüf Stiftung (GRS-057/14) and the Novartis Research Foundation (to N.H.T.); and NIH grants NCI R01CA214608 and R01CA218278 and a Mark Foundation Emerging Leader Award (to E.S.F.). M.S. received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie grant agreement no. 702642; Z.K. was supported by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie grant agreement no. 765445; G.P. was supported by the Human Frontier Science Program (HFSP Long-Term Fellowship LT000210/2014) and the European Molecular Biology Organization (EMBO Advanced Fellowship aALTF 761-2016); A.S.S. was supported by a DF/HCC K12 grant, a Conquer Cancer Foundation Young Investigator Award and an award from the Wong Family Foundation; R.S.S. was supported by an Intermediate Fellowship from the Kay Kendall Leukaemia Fund and an Advanced Clinician Scientist Fellowship from Cancer Research UK; and S.M.C. received funding from grants KL2 TR002542 and K08 CA230220.

Author contributions M.S. designed and performed functional genomics studies with the help of J.K., R.S.S. and E.C.F.; Z.K. and G.P. designed and carried out structural, biochemical and biophysical studies with the help of D.S.; M.S., Y.-D.L., M.M. and Q.L.S. designed and performed cell validation experiments with the help of A.S.S., J.A.G. and M.J.; K.A.D. performed the mass spectrometry experiments; M.S., L.R.W. and S.M.C. performed bioinformatic PRISM analysis; R.B. and G.P. performed structure refinement with the help of Z.K.; D.G., C.S., S.F., T.R.G., E.S.F., N.H.T. and B.L.E. supervised the project; and Z.K., G.P., M.S., B.L.E. and N.H.T. wrote the manuscript with input from all authors.

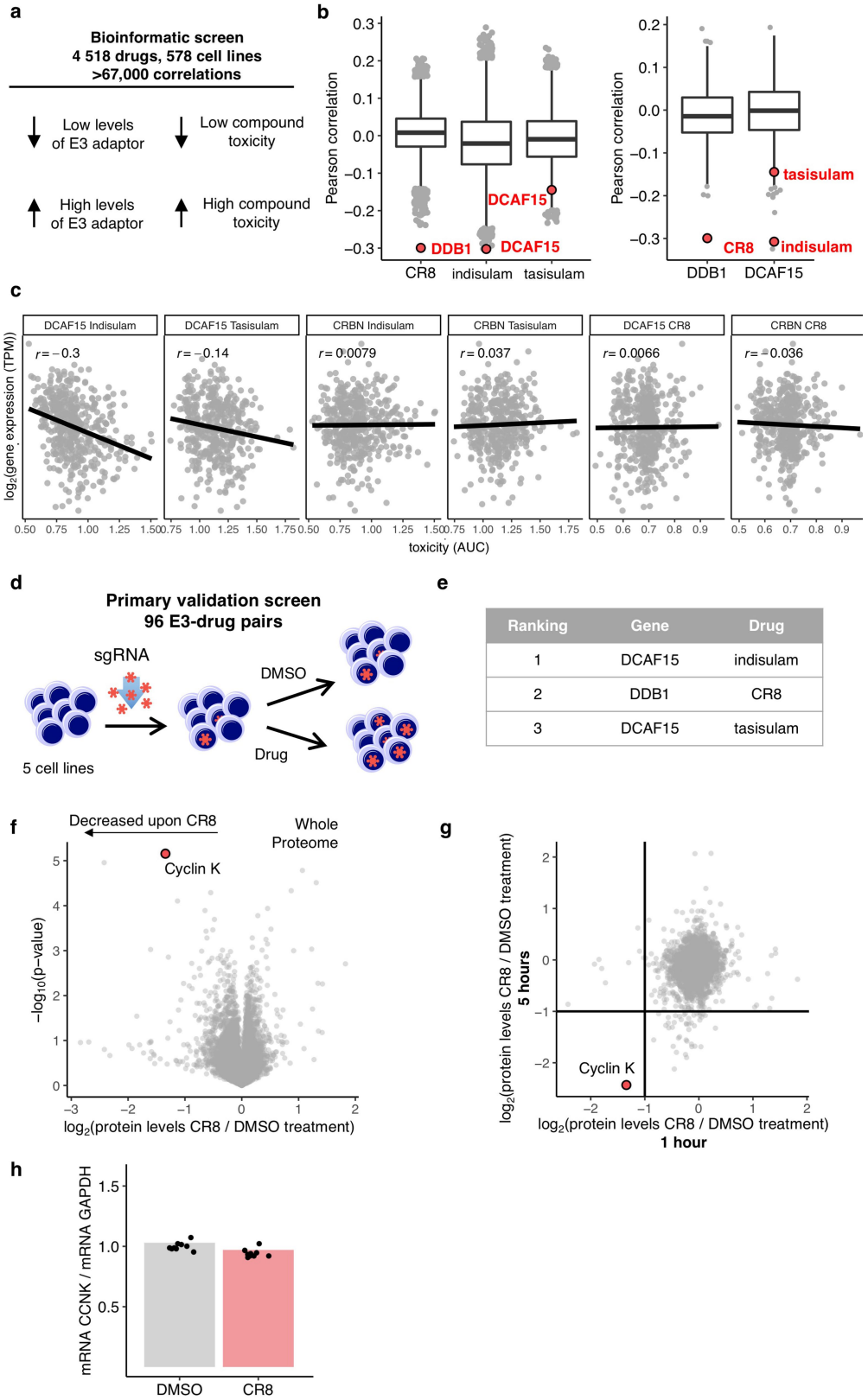
Competing interests B.L.E. has received research funding from Celgene and Deerfield and consulting fees from GRAIL. He serves on the scientific advisory boards for and holds equity in Skyhawk Therapeutics and Exo Therapeutics. E.S.F. is a founder and/or member of the scientific advisory board and equity holder of C4 Therapeutics and Civetta Therapeutics, and is a consultant to Novartis, AbbVie and Pfizer. N.H.T. receives funding from the Novartis Research Foundation and is a member of the scientific advisory board of Monte Rosa Therapeutics. The Fischer laboratory receives or has received research funding from Novartis, Deerfield and Astellas. S.F. has had a consulting or advisory role, received honoraria, research funding and/or travel/accommodation expenses funding from the following for-profit companies: Bayer, Roche, Amgen, Eli Lilly, PharmaMar, AstraZeneca and Pfizer. R.B. is now an employee of Monte Rosa Therapeutics. S.M.C. and T.R.G. receive research funding from Bayer HealthCare. T.R.G. was formerly a consultant and equity holder in Foundation Medicine (acquired by Roche). T.R.G. also is a consultant to GlaxoSmithKline and is a founder of Sherlock Biosciences.

Additional information

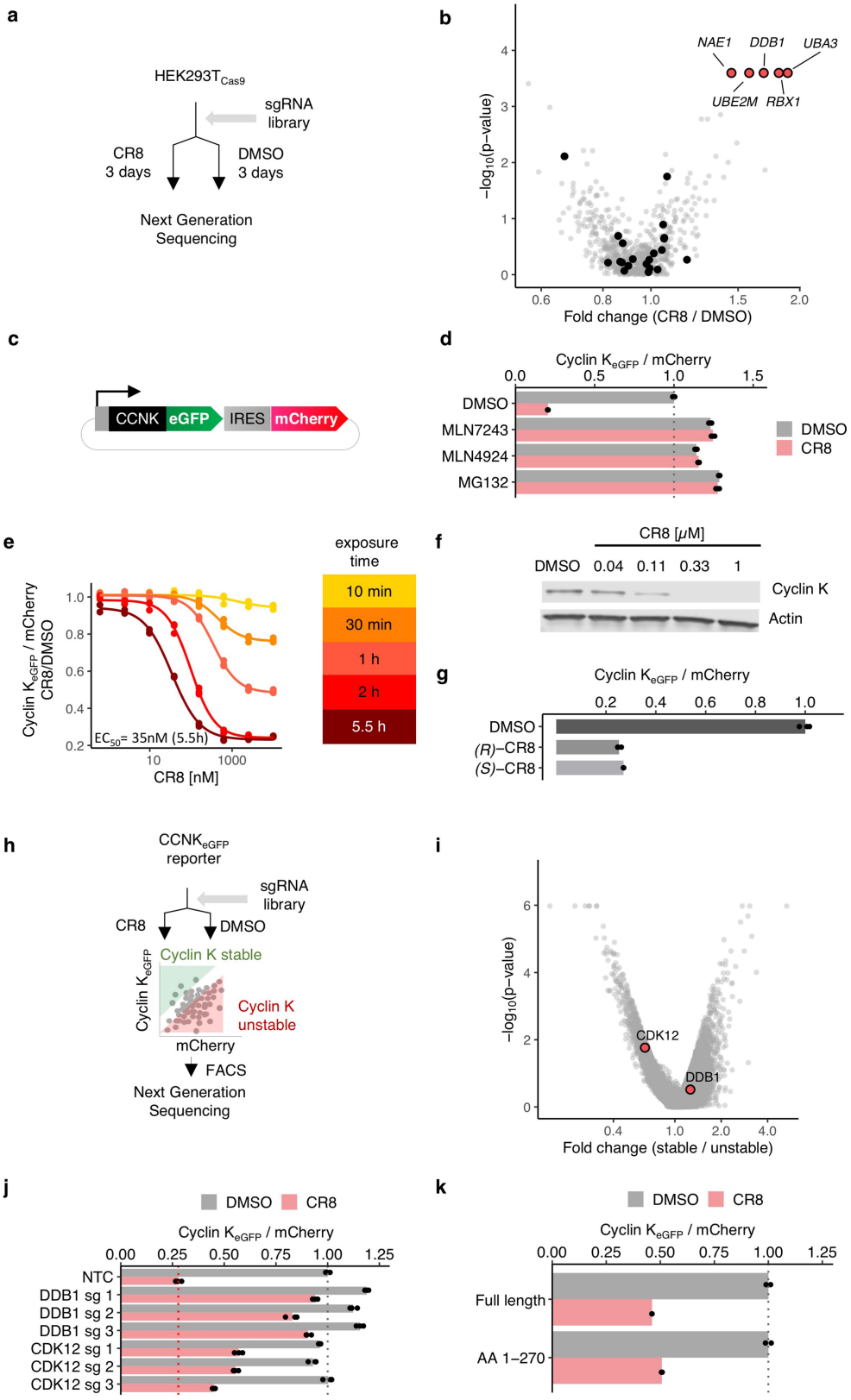
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2374-x.

Correspondence and requests for materials should be addressed to N.H.T. or B.L.E.

Peer review information *Nature* thanks Ivan Dikic, Frank Sicheri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

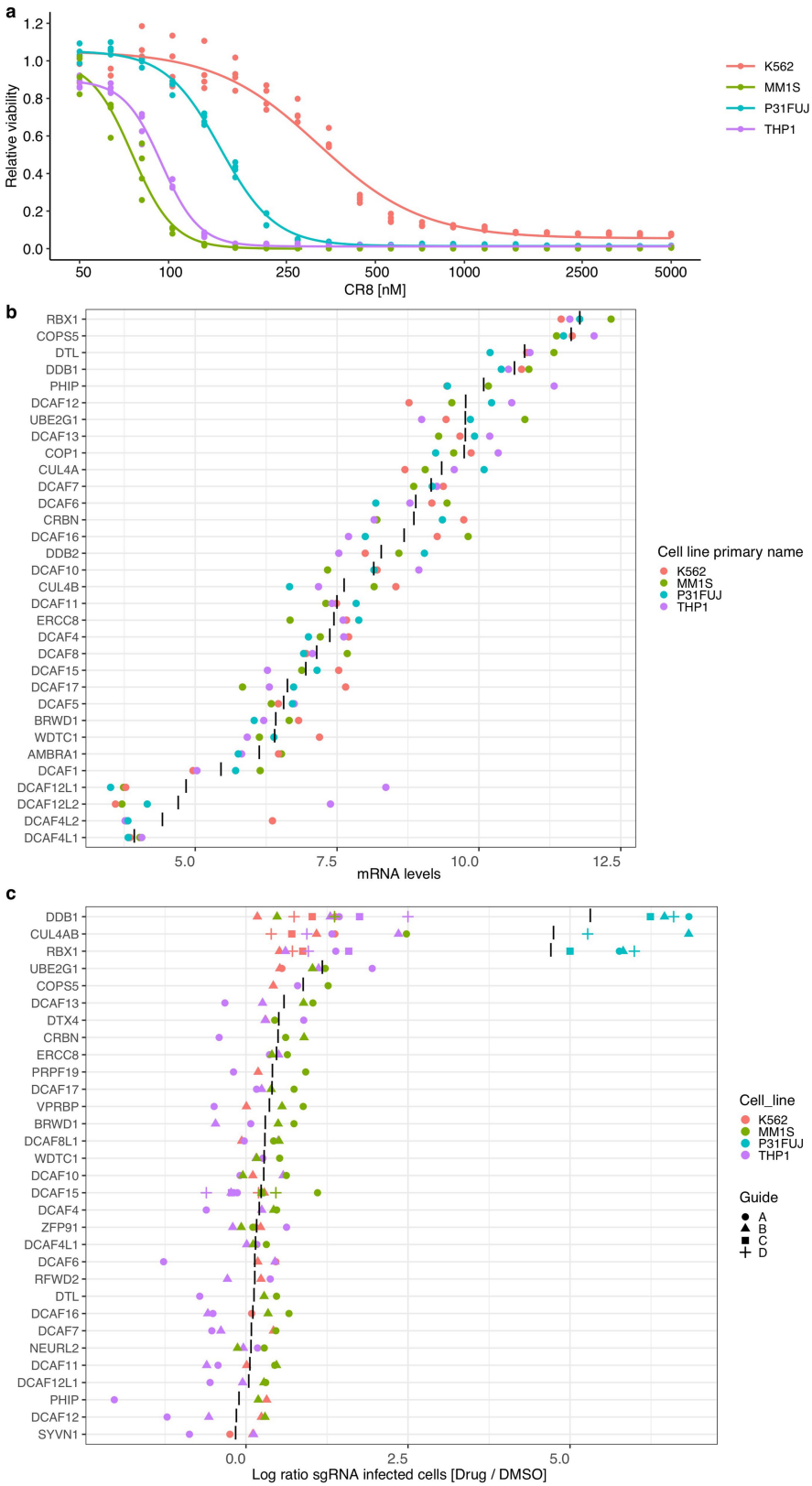


Extended Data Fig. 1 | CR8-induced degradation of cyclin K correlates with DDB1 expression. a, Schematic of the bioinformatic screen for drug–E3 pairs. **b**, Box plot (centre, median; box, interquartile range (IQR); whiskers, 1.5 × IQR; points, outliers) for correlations between gene expression and drug sensitivity (CR8 $n = 19{,}110$; indisulam and tasisulam $n = 19{,}109$; *DDB1* and *DCAF15* $n = 1{,}618$). **c**, Example Pearson correlation of selected drug–E3 pairs: positive controls (indisulam and *DCAF15*; tasisulam and *DCAF15*) and no correlation controls (others) (indisulam $n = 452$; tasisulam $n = 418$; CR8 $n = 471$), area under the curve (AUC). **d**, Schematic of flow-based primary validation screen. **e**, Top three hits from the primary validation screen in five cell lines, performed according to the schematic in **d**. **f**, Whole-proteome quantification of MOLT-4 cells treated with 1 µM CR8 ($n = 1$) or DMSO ($n = 3$) for 1 h (two-sided moderated *t*-test, $n = 3$). **g**, Log_2(protein levels CR8 treatment/protein levels DMSO treatment) in whole-proteome quantification after 1 h and 5 h of exposure to CR8 plotted against each other. **h**, mRNA levels quantified by qPCR in HEK293T-Cas9 cells after treatment with 1 µM CR8 for 2 h. Bars represent the mean ($n = 9$).

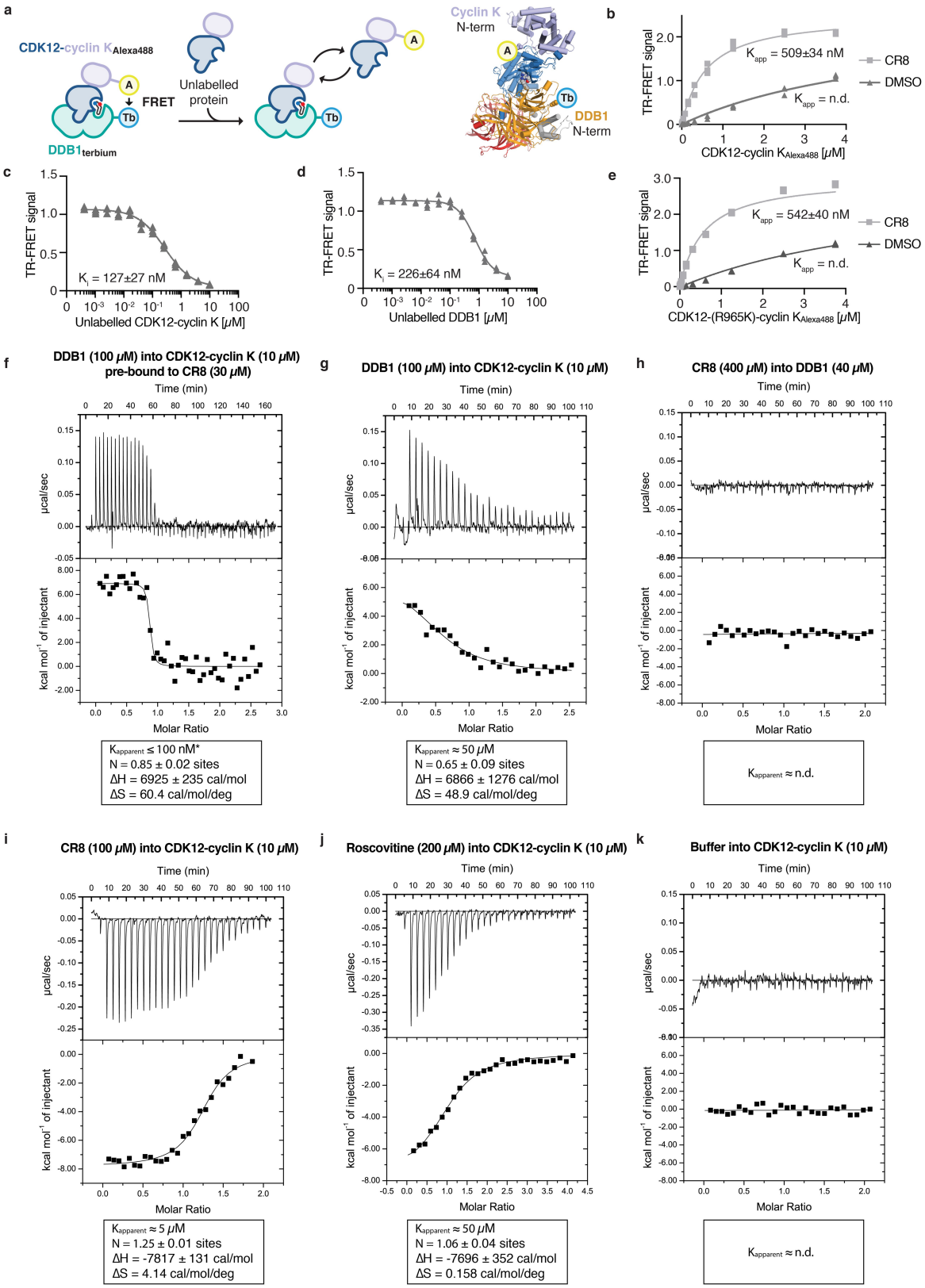


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | CDK12 is required for CR8-induced cyclin K degradation. a, Schematic of the genome-wide CRISPR–Cas9 resistance screen. **b**, Bison CRISPR–Cas9 CR8 resistence screen. Guide counts were collapsed to gene level ($n = 4$ guides per gene; two-sided empirical rank-sum test statistics). **c**, Schematic of the cyclin K (*CCNK*) stability reporter. IRES, internal ribosome entry site. **d**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells that were pretreated with 0.5 μM MLN7243, 1 μM MLN4924 or 10 μM MG132 for 4 h and then treated with 1 μM CR8 for 2 h ($n = 3$). **e**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells treated with CR8 ($n = 3$). **f**, Immunoblots of cyclin K degradation in HEK293T-Cas9 cells treated with CR8 for 2 h ($n = 2$). **g**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells treated with 1 μM of the indicated compounds for 2 h ($n = 3$). **h**, Schematic of the genome-wide CRISPR–Cas9 reporter screen for cyclin K stability. **i**, Genome-wide CRISPR–Cas9 reporter screen for cyclin K_{eGFP} stability with DMSO treatment in HEK293T-Cas9 cells. Guide counts were collapsed to gene level ($n = 4$ guides per gene; two-sided empirical rank-sum test statistics). **j**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells after treatment with 1μM CR8 for 2 h ($n = 3$). **k**, Flow analysis of full-length cyclin K_{eGFP} or cyclin K_{eGFP}(1–270) in HEK293T-Cas9 cells after treatment with 1 μM CR8 for 2 h ($n = 3$). Bars represent the mean in **d**, **g**, **j**, **k**.

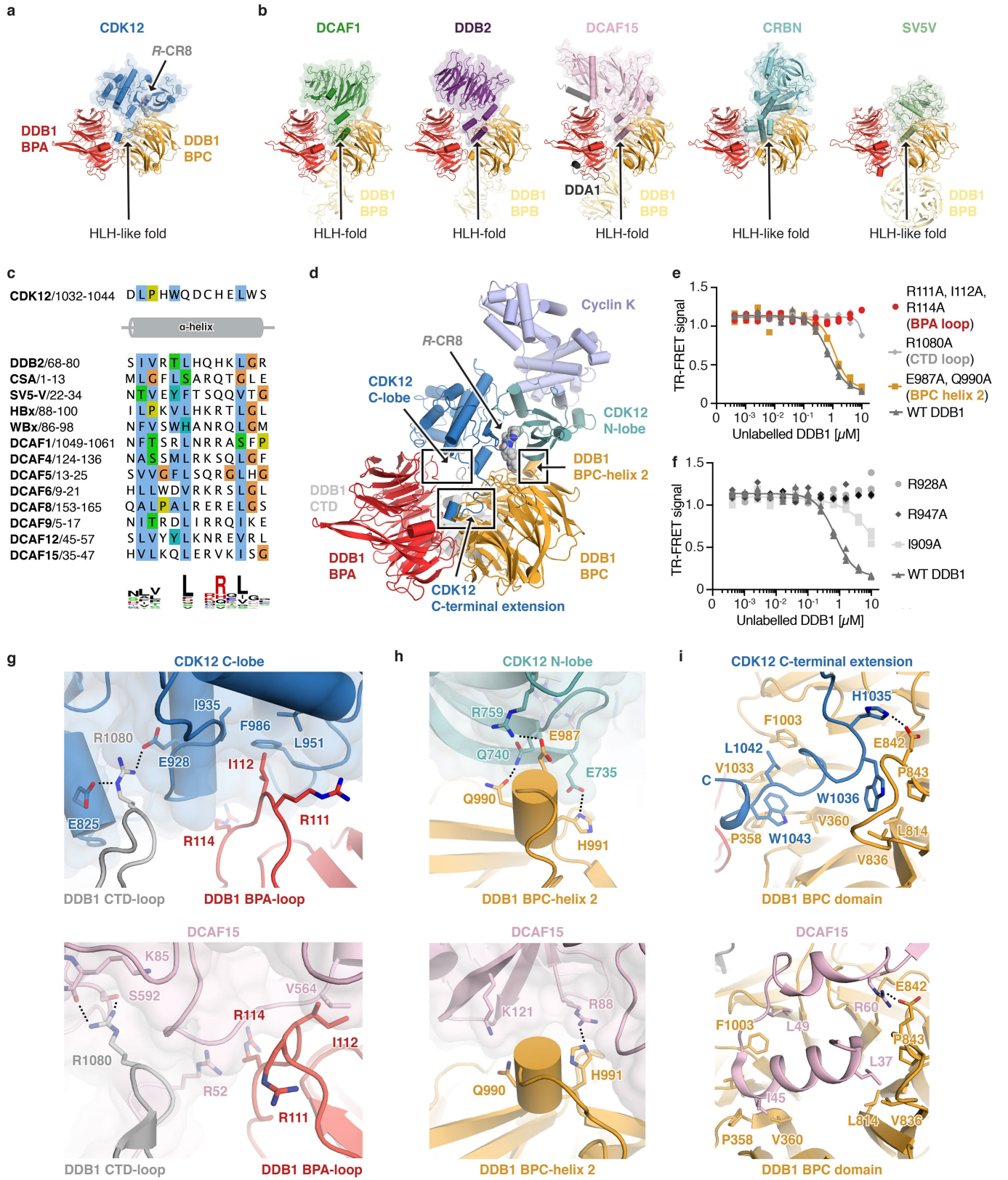


Extended Data Fig. 3 | CR8-induced degradation of cyclin K is not dependent on a canonical DCAF substrate receptor. a, Drug sensitivity of K562-Cas9, P31FUJ-Cas9, THP1-Cas9 and MM1S-Cas9 cells that were treated with CR8 for three days ($n = 3$). **b**, mRNA levels of genes in the DCAF library. The black vertical lines represent the mean ($n = 4$). **c**, Flow analysis of K562-Cas9, P31FUJ-Cas9, THP1-Cas9 and MM1S-Cas9 cells expressing sgRNAs and a BFP marker after a three-day treatment with 1 μM CR8. The black vertical lines represent the mean ($n > 2$).

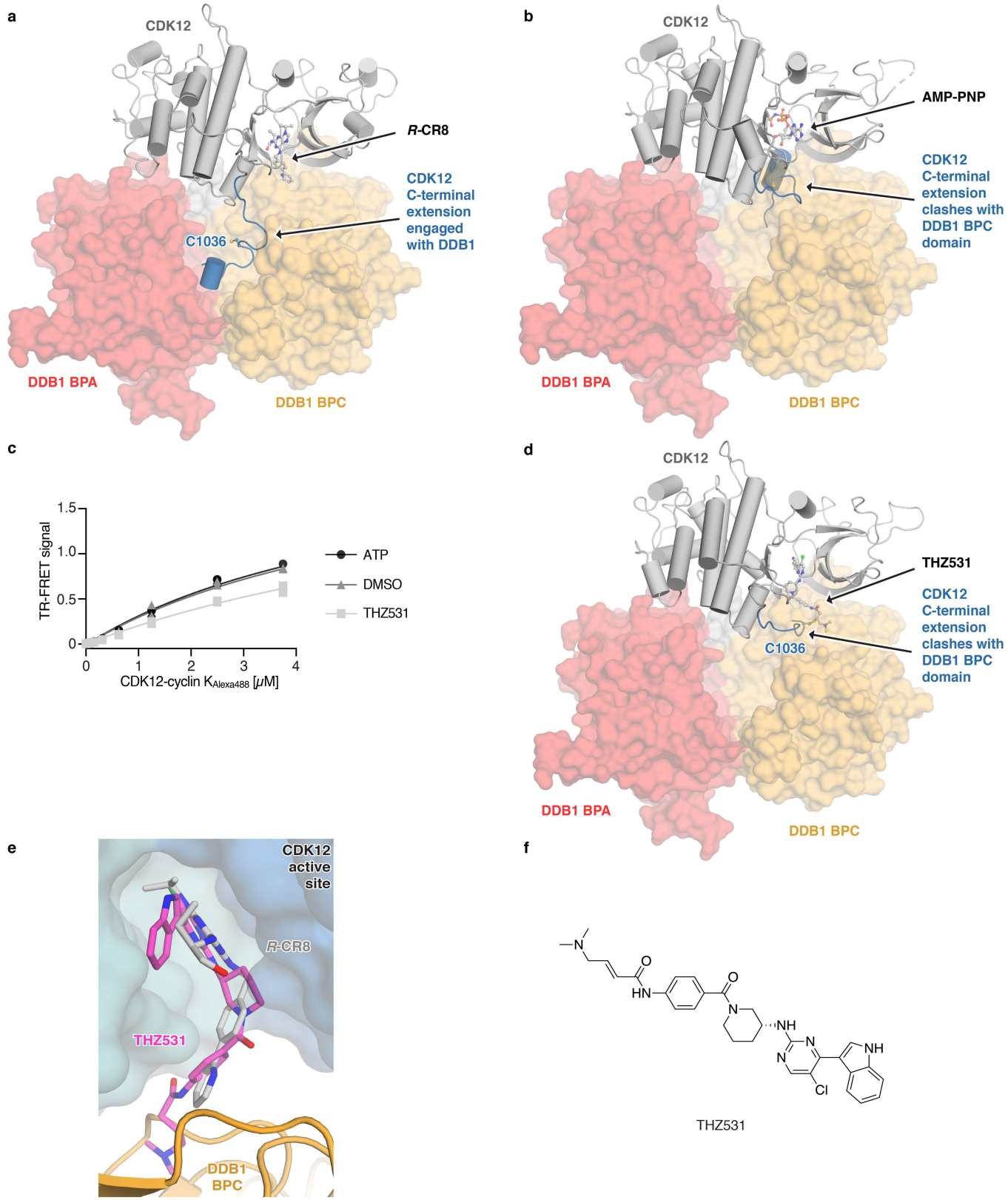


Extended Data Fig. 4 | See next page for caption.

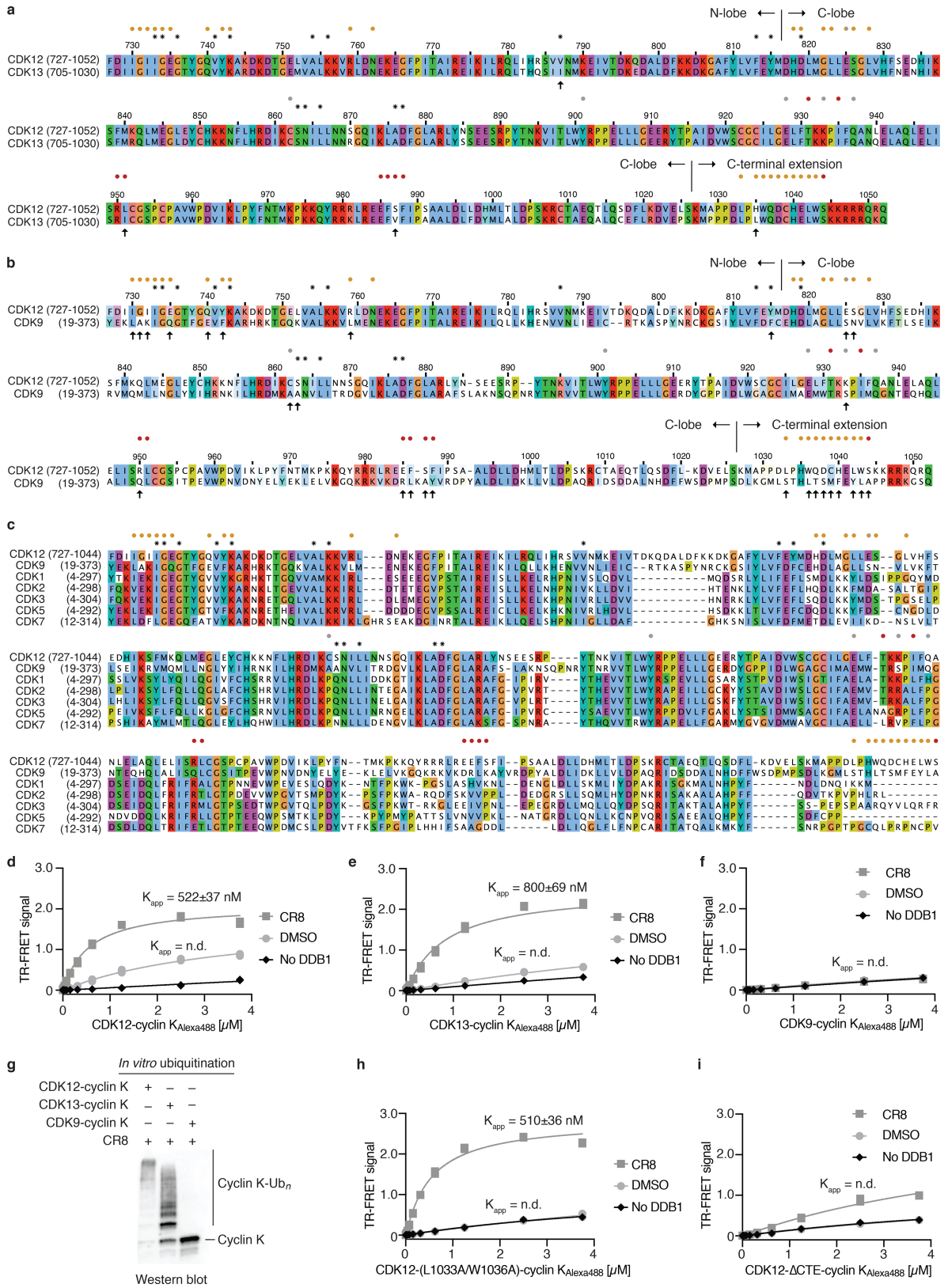
Extended Data Fig. 4 | Characterization of DDB1–CDK12–cyclin K complex formation. a, Schematic of the TR-FRET set-up. Positions of FRET donor (terbium-coupled streptavidin (Tb)) and acceptor (Alexa488–SpyCatcher (A)) are indicated in the structural model. **b**, Titration of CDK12–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). **c**, Counter-titration of unlabelled wild-type CDK12–cyclin K (0–10 μM) into 50 nM $DDB1_{terbium}$, 500 nM CDK12–cyclin $K_{Alexa488}$ and 12.5 μM CR8 ($n = 3$). **d**, Counter-titration of unlabelled wild-type DDB1 (0–10 μM) into 50 nM $DDB1_{terbium}$, 500 nM CDK12–cyclin $K_{Alexa488}$ and 1 μM CR8 ($n = 3$). **e**, Titration of CDK12(R965K)–cyclin $K_{Alexa488}$ (wild-type sequence of the canonical isoform of CDK12; 0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). The CDK12(K965R) variant, which was used throughout our in vitro studies (see Methods), shows a binding affinity indistinguishable from that of the canonical isoform of CDK12 (residue distal from the interface with DDB1 and cyclin K). **f**–**k**, ITC experiments. Specifications of the titrations are given in the panels. N.d., not determined. ΔS, entropy of binding. K_d values are referred to as $K_{apparent}$ because not all data could be confidently fitted with binding curves. An asterisk marking the approximate $K_{apparent}$ value in **f** denotes that the binding affinity was too high to allow precise affinity determination. $n = 2$ (**f**, **g**); $n = 1$ (**h**, **k**); $n = 3$ (**i**, **j**); additional replicates are provided in the Supplementary Information.



Extended Data Fig. 5 | CDK12 contacts residues on DDB1 that are otherwise involved in DCAF binding. a, Structure of the DDB1(ΔBPB)–(*R*)-CR8–CDK12 complex. The CDK12 C-terminal extension binds a cleft between the DDB1 BPA and BPC domains (arrow) and adopts an helix–loop–helix (HLH)-like fold. **b**, Diverse DCAFs bind DDB1 through HLH or HLH-like folds. **c**, Sequence alignment of identically positioned helices of different HLH domains. **d**, Overview of protein–protein interaction hotspots. **e**, Counter-titration of unlabelled wild-type or mutant DDB1 (0–10 μM) into the preassembled $DDB1_{terbium}$–CR8–CDK12–cyclin $K_{Alexa488}$ complex ($n = 3$). **f**, Counter-titration of unlabelled wild-type or mutant DDB1 (0–10 μM) into the preassembled $DDB1_{terbium}$–CR8–CDK12–cyclin $K_{Alexa488}$ complex ($n = 3$). **g**–**i**, Close-up views of DDB1 residues contacted by CDK12 (top) that are otherwise involved in DCAF binding (bottom).

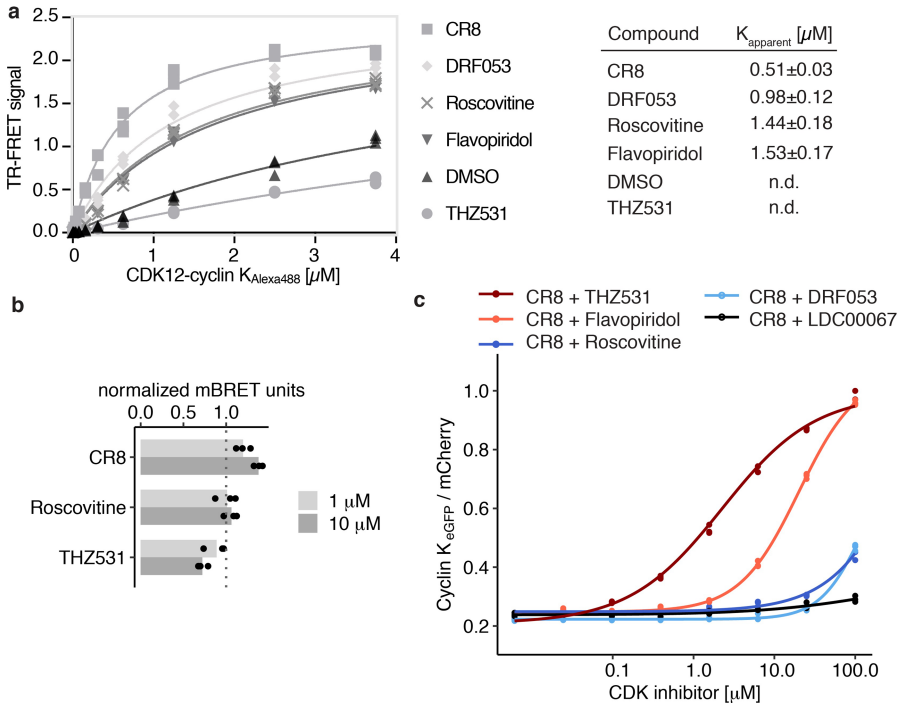


Extended Data Fig. 6 | The CDK12 C-terminal extension adopts different conformations. a, Conformation of the C-terminal extension in the structure of the DDB1(ΔBPB)–CR8–CDK12 complex. **b**, Structure of CDK12 bound to adenylyl-imidodiphosphate (AMP-PNP) (PDB entry 4CXA) superimposed onto CDK12 in the DDB1(ΔBPB)–CR8–CDK12 complex. **c**, Titration of CDK12–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ in the presence of 10 μM THZ531, ATP or DMSO ($n = 3$). **d**, Structure of CDK12 bound to THZ531 (PDB entry 5ACB) superimposed onto CDK12 in the DDB1(ΔBPB)–CR8–CDK12 complex. **e**, THZ531-binding pose in the active site of CDK12 as in **d**. **f**, Chemical structure of THZ531.

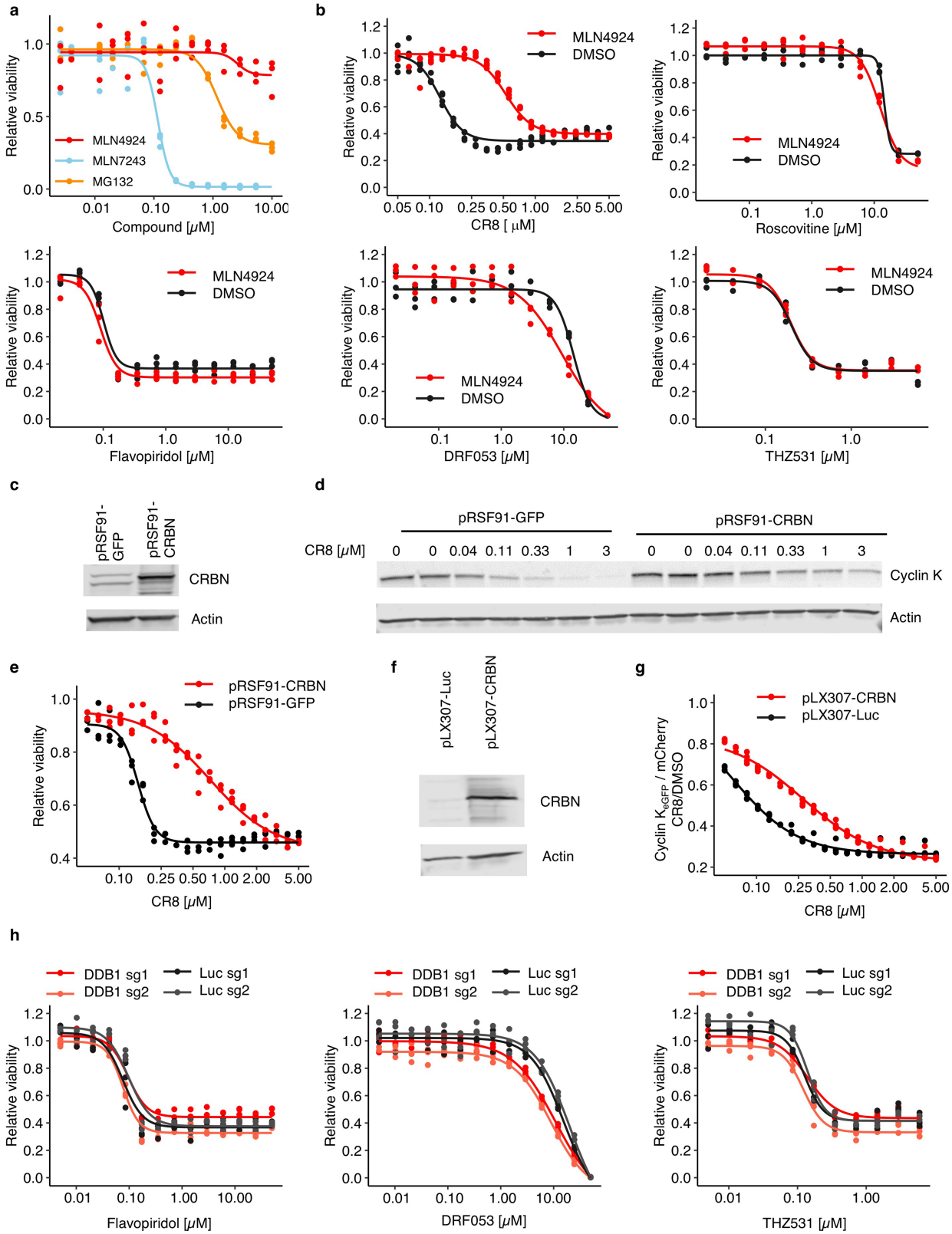


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Differences between CDK12 and other CDKs highlight the residues involved in CR8-induced recruitment of DDB1. **a**, Sequence alignment of CDK12 and CDK13. **b**, Sequence alignment of CDK12 and CDK9. **c**, Multiple sequence alignment of different human CDKs. In **a**–**c**, asterisks denote contacts with CR8 and circles indicate contacts with DDB1 (coloured according to DDB1 domains; see Fig. 2). Arrows mark differences at the DDB1–CR8–CDK interface. **d**, Titration of CDK12–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). 'No DDB1' only contains terbium-coupled streptavidin and shows concentration-dependent fluorophore effects. **e**, Titration of CDK13–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). **f**, Titration of CDK9–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). **g**, $CUL4_{NEDD8}$–RBX1–DDB1 in vitro ubiquitination of cyclin K bound to CDK12, CDK13 or CDK9 ($n = 2$). **h**, Titration of CDK12(L1033A/W1036A)–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). **i**, Titration of CDK12(ΔCTE)–cyclin $K_{Alexa488}$ (0–3.75 μM) into 50 nM $DDB1_{terbium}$ and 10 μM CR8 or DMSO ($n = 3$). CDK12(ΔCTE) is a truncated version of CDK12 (amino acids 713–1032).



Extended Data Fig. 8 | CDK inhibitors block the CR8-induced degradation of cyclin K. a, Titration of CDK12–cyclin $K_{Alexa488}$ into $DDB1_{terbium}$ in the presence of the indicated compounds (all 10 μM) ($n = 3$). **b**, NanoBRET of HEK293T cells transfected with NanoLuc-labelled CDK12(713–1052) and HaloTag-labelled DDB1(ΔBPB) constructs and treated with the indicated compounds for 2 h. Bars represent the mean ($n = 3$). **c**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells treated with 1 μM CR8 and competitive CDK inhibitor ($n = 3$).



Extended Data Fig. 9 | Cytotoxicity of CR8 analogues does not depend on CRL4 components. a, Drug sensitivity of HEK293T-Cas9 cells that were treated with different inhibitors for three days ($n = 3$). **b**, Drug sensitivity of HEK293T-Cas9 cells that were treated with 100 nM MLN4924 or DMSO in combination with the indicated compound for three days ($n = 3$). **c**, Immunoblots of HEK293T-Cas9 cells transfected with control (pRSF91-GFP) or CRBN overexpression (pRSF91-CRBN) vectors ($n = 2$). **d**, Immunoblots of HEK293T-Cas9 cells expressing pRSF91-GFP or pRSF91-CRBN and treated with CR8 for three days ($n = 2$). **e**, Drug sensitivity of HEK293T-Cas9 cells expressing pRSF91-GFP or pRSF91-CRBN and treated with CR8 for three days ($n = 3$). **f**, Immunoblots of HEK293T-Cas9 cells transfected with control (pLX307-Luc) or CRBN overexpression (pLX307-CRBN) vectors ($n = 2$). **g**, Flow analysis of cyclin K_{eGFP} degradation in HEK293T-Cas9 cells expressing pLX307-Luc or pLX307-CRBN and treated with CR8 for 2 h ($n = 3$). **h**, Drug sensitivity of HEK293T-Cas9 cells expressing sgRNAs targeting *DDB1* or luciferase and treated with the indicated inhibitor for three days ($n = 3$).

Extended Data Table 1 | Data collection and refinement statistics

	DDB1$_{\Delta BPB}$-CR8-CDK12$_{713-1052}$-cycK$_{1-267}$
Data collection*	
Space group	$P3_121$
Cell dimensions	
a, *b*, *c* (Å)	250.75, 250.75, 217.92
α, β, γ (°)	90, 90, 120
Resolution (Å)	54–3.46 (3.63–3.46) †
R_{meas}	0.318 (>4.00)
$I / \sigma I$	7.2 (0.9)
Completeness (%)	95.1 (68.3)‡
Redundancy	12.0 (11.6)
Refinement	
Resolution (Å)	54–3.46
No. reflections	89,183
R_{work} / R_{free}	0.1934 / 0.220
No. atoms	
Protein	33,781
Ligand/ion	96
Water	0
B-factors	
Protein	59.9
Ligand/ion	39.6
Water	n/a
R.m.s. deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.01

*Data collected from a single crystal. †Values in parentheses are for the highest-resolution shell. ‡From STARANISO[42] assuming a local weighted $CC_{1/2}$ = 0.3 resolution cut-off.

nature research

Corresponding author(s): Nicolas H. Thoma and Benjamin L. Ebert

Last updated by author(s): Apr 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Proteins were identified and quantified using Proteome Discoverer 2.2 (Thermo Fisher Scientific): RRID:SCR_014477 QPCR signal was quantified with CFX96 Real Time system (Bio-Rad). Western blot data were imaged by Odyssey Imaging System, Image Studio (Li-Cor). The luminescent BRET signal was acquired with FilterMax F5 plate reader (Molecular Devices) or EnVision Multilabel Plate Reader (Perkin Elmer). The luminescent CTG signal was acquired with EnVision Multilabel Plate Reader (Perkin Elmer) or CLARIOstar Plus, MARS 3.4 (BMG LabTech). Flow data were collected with BD FACSDiva 8.0 (BD Biosciences) or CytExpert Software (Beckman Coulter Life Sciences). TR-FRET data collection was performed using a Pherastar FS (BMG).
Data analysis	Crystallographic data processing and refinement was done using XDS, DIALS (2.0), AIMLESS (CCP4 suite 7.0), STARANISO, PHASER, COOT (0.9), phenix (1.17.1-3660), BUSTER, eLBOW, PyMOL (2.3.2), MOLPROBITY/PDB-REDO. Fluorescence polarization fits were done using Prism7 (Graphpad). Bioinformatic, CRISPR screen data analyses and data visualization were done using R programming 3.5.1 and RStudio (1.1.453) with the following packages: tidyverse (1.2.1), ggrepel (0.8.1), GGally (1.4.0), dr4pl (1.1.11), ShortReads (Bioconductor) and Statistical Analysis Limma Package (Bioconductor). Custom R scripts were used to analyze the data, which are attached as Supplementary Code. Flow data were analyzed with FlowJo 10 (BD).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Structural data is deposited in the PDB under the accession code 6TD3. Gene expression data and drug sensitivity data are available through the DepMap portal (https://depmap.org). Drug sensitivity data are also archived via Figshare (doi:10.6084/m9.figshare.9393293). For the manuscript analysis, we used a provisional data set, however, all the findings can be reproduced from the final release of the data. Proteome quantification data are available in the PRIDE repository (PXD016187 and PXD016188). Uncropped western blot gels can be found in the Supplementary Information.

Figures that have associated raw data:

Extended Data Fig. 1e is associated with Primary data for validation of 96 E3 gene-compound pairs. Fig. 1c and Extended Data Fig. 1f, g is associated with Supplementary Data: Proteome quantification using tandem mass tag spectrometry data. Fig1 f,g, Extended Data Fig. 2b,i is associated with Supplementary Data: Functional genomics data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The sample size for the number of replicates (n) for each experiment is provided in the figure captions.
Data exclusions	No data was excluded from the analysis. For functional genomics experiments, data processing, which includes filtering, is described in Methods section and can be re-analyzed with attached R-programming scripts.
Replication	Data presented in the figure represent technical replicates, however drug treatment or perturbation was performed independently for each replicate. Replications were consistent across multiple experiments performed on different days.
Randomization	No randomization was performed since no animal-based experiments were performed.
Blinding	Investigators were not blinded during data collection or analysis. However, controls and samples were analyzed in exactly the same way using the same computational pipeline.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-cycK (Bethyl Laboratories, A301-939A for full length cycK, 1:2,000, Lot#6), anti-cycK (abcam, ab251652, for cycK1-267, 1:2000), anti-beta-actin (Cell Signaling, #3700, 1:5000, lot#17), anti-CRBN (Sigma prestige, HPA045910, 1:1000, lot#Q103829), anti-mouse 800CW (LI-COR Biosciences, 926-32211, 1:10 000, lot#C90917-25), anti-rabbit 680LT (LI-COR Biosciences, 925-68021, 1:10 000, lot#C90501-05), and anti-rabbit IgG antibodies (abcam, ab6721) were used in this study.

Validation	All antibodies are commercially available. Anti-cycK (Bethyl Laboratories), anti-beta-actin, and anti-CRBN were validated for immunoblotting by their manufacturer. Anti-cycK (Abcam) were validated only for immunocytochemistry and immunofluorescence by their manufacturer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The human HEK293T cell lines were provided by the Genetic Perturbation Platform, Broad Institute and K562-Cas9 , THP1-Cas9, P31FUJ-Cas9 cell lines were provided by Zuzana Tothova (Broad Institute) and HEK293T-Cas9 and MM1S-Cas9 were previously published (see Methods section). Sf9 (purchased from Thermo Fischer Scientific Cat# 11496-015) and Hi5 cells (Tni cells, purchased from Expression Systems Cat# 94-002F) were also used in this study. MOLT-4 were purchased from ATCC.
Authentication	HEK293T, K562-Cas9, THP1-Cas9, P31FUJ-Cas9, HEK293T-Cas9,MM1S-Cas9, and MOLT-4 cell lines were authenticated by STR profiling. Sf9 and Hi5 cells were authenticated by the vendor.
Mycoplasma contamination	Mycoplasma negative.
Commonly misidentified lines (See ICLAC register)	None of commonly misidentified lines were used in this study.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Adherent cells were trypsinized, collected, and the cell pellets resuspended in PBS. Suspension cells were washed with PBS or directly subjected to analysis without fixation.
Instrument	Cytoflex LX (Beckman), MA900 Cell Sorter (Sony), Fortessa FACS (BD Biosciences), FACSCanto (BD Biosciences).
Software	BD FACSDiva 8.0 (BD Biosciences), CytExpert Software (Beckman), FlowJo 10 (BD).
Cell population abundance	Round cells (population with forward and side scatter properties consistent with the alive, non-treated cell line) were usually > 50% in most measurements (rarely lower due to drug toxicity), singlets were > 90%. For reporter assays, mCherry positive cells > 50%.
Gating strategy	Cells were first gated for live cells based on forward and side scatter. Single cells were discriminated based on the area vs. height of the side scatter. Finally, reporter positive cells were gated based on the mCherry expression.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Nucleolar RNA polymerase II drives ribosome biogenesis

https://doi.org/10.1038/s41586-020-2497-0

Received: 19 December 2018

Accepted: 21 April 2020

Published online: 15 July 2020

Check for updates

Karan J. Abraham[1,11], Negin Khosraviani[1,11], Janet N. Y. Chan[1], Aparna Gorthi[2], Anas Samman[1], Dorothy Y. Zhao[3,4], Miling Wang[5], Michael Bokros[5], Elva Vidya[1], Lauren A. Ostrowski[1], Roxanne Oshidari[1], Violena Pietrobon[1], Parasvi S. Patel[6], Arash Algouneh[1,6], Rajat Singhania[6], Yupeng Liu[1], V. Talya Yerlici[1], Daniel D. De Carvalho[6], Michael Ohh[1,7], Brendan C. Dickson[1,8], Razq Hakem[6], Jack F. Greenblatt[3,4], Stephen Lee[5], Alexander J. R. Bishop[2,9] & Karim Mekhail[1,10]✉

Proteins are manufactured by ribosomes—macromolecular complexes of protein and RNA molecules that are assembled within major nuclear compartments called nucleoli[1,2]. Existing models suggest that RNA polymerases I and III (Pol I and Pol III) are the only enzymes that directly mediate the expression of the ribosomal RNA (rRNA) components of ribosomes. Here we show, however, that RNA polymerase II (Pol II) inside human nucleoli operates near genes encoding rRNAs to drive their expression. Pol II, assisted by the neurodegeneration-associated enzyme senataxin, generates a shield comprising triplex nucleic acid structures known as R-loops at intergenic spacers flanking nucleolar rRNA genes. The shield prevents Pol I from producing sense intergenic noncoding RNAs (sincRNAs) that can disrupt nucleolar organization and rRNA expression. These disruptive sincRNAs can be unleashed by Pol II inhibition, senataxin loss, Ewing sarcoma or locus-associated R-loop repression through an experimental system involving the proteins RNaseH1, eGFP and dCas9 (which we refer to as 'red laser'). We reveal a nucleolar Pol-II-dependent mechanism that drives ribosome biogenesis, identify disease-associated disruption of nucleoli by noncoding RNAs, and establish locus-targeted R-loop modulation. Our findings revise theories of labour division between the major RNA polymerases, and identify nucleolar Pol II as a major factor in protein synthesis and nuclear organization, with potential implications for health and disease.

Various proteins self-organize via liquid–liquid phase separation (LLPS) into nucleolar subdomains, which are needed for highly stereotyped ribosome assembly[1,2]. At fibrillar centres in the heart of mammalian nucleoli, the major rRNA molecules needed to assemble ribosomes are generated by Pol-I-dependent transcription of rRNA genes within ribosomal DNA (rDNA) repeats[1,3]. Within rDNA, rRNA genes are separated by large intergenic spacers (IGSs) (Extended Data Fig. 1a). At nucleolar rRNA genes, Pol I synthesizes precursor rRNAs (pre-rRNAs) that are processed into mature 28S, 18S and 5.8S rRNA molecules as they migrate to the granular component at the nucleolar periphery. Outside nucleoli, Pol III synthesizes 5S rRNA molecules that are targeted to nucleoli for processing. Mature rRNAs are packaged into 40S and 60S ribosomal subunits for export to the cytoplasm. Traditionally, the nucleolar Pol I and nucleoplasmic Pol III are viewed as the sole mammalian RNA polymerases that directly mediate housekeeping ribosome biogenesis. Interestingly, in the budding yeast *Saccharomyces cerevisiae*, Pol II is physically enriched at rDNA IGSs, but this phenomenon is deleterious because it drives ageing without affecting rRNA expression[3–5]. It is unclear whether nucleolar Pol II exists in higher organisms or directly promotes ribosome biogenesis in any species.

Active Pol II at rDNA IGSs

To investigate whether Pol II exists within human nucleoli, we first used immunofluorescence coupled to super-resolution microscopy. Within nucleoli, which were outlined by nucleophosmin (NPM), we observed foci corresponding to active Pol II phosphorylated on serine 2 (pS2) (Fig. 1a and Extended Data Fig. 1b, c). Chromatin immunoprecipitation (ChIP) showed that pS2 and another active form of Pol II, phosphorylated on serine 5 (pS5), were enriched across rDNA, with the highest levels—at IGS28 and IGS38—being comparable to those at known Pol-II-transcribed loci (Fig. 1b and Extended Data Fig. 1a, d–f).

[1]Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [2]Greehey Children's Cancer Research Institute, Department of Cell Systems and Anatomy, University of Texas Health at San Antonio, San Antonio, TX, USA. [3]Department of Molecular Genetics, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [4]Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [5]Sylvester Comprehensive Cancer Center, Department of Biochemistry and Molecular Biology, Miller School of Medicine, University of Miami, Miami, FL, USA. [6]Princess Margaret Cancer Research Centre, University Health Network, and Department of Medical Biophysics, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [7]Department of Biochemistry, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [8]Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [9]Mays Cancer Center, University of Texas Health at San Antonio, San Antonio, TX, USA. [10]Canada Research Chairs Program, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. [11]These authors contributed equally: Karan J. Abraham, Negin Khosraviani. ✉e-mail: karim.mekhail@utoronto.ca

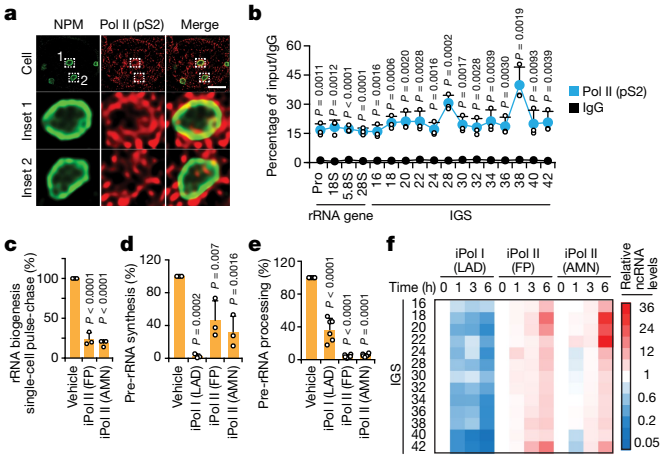


Fig. 1 | Pol I and Pol II localize to rDNA IGSs and compete to modulate IGS ncRNA levels. a, Representative immunofluorescence and super-resolution microscopy images showing the localization of pS2 Pol II within NPM-delineated nucleoli. Scale bar, 5 µm. **b**, Enrichment of pS2 Pol II across rDNA as revealed by chromatin immunoprecipitation (ChIP). Enrichments, accounting for typical background fluctuations across repetitive DNA loci, were calculated as (percentage of input/IgG) = (percentage of input for protein immunoprecipitation)/(percentage of input for mock IgG immunoprecipitation). **c**, Effect of a 3-hour Pol II inhibition (iPol II) using flavopiridol (FP) or α-amanitin (AMN) on rRNA biogenesis as measured in live single-cell pulse-chase assays using 5-fluorouracil (FU)-labelled RNA. **d**, **e**, Cell-population-based RNA pulse-chase assays were used to assess pre-rRNA synthesis (**d**) and processing (**e**) following a 3-hour inhibition of Pol I or Pol II (iPol I/II; low-dose actinomycin-D, LAD). **f**, Pol I promotes, and Pol II represses, IGS ncRNAs, as shown by reverse transcription with quantitative polymerase chain reaction (RT–qPCR). **a**–**f**, Experiments carried out with HEK293T cells; data shown as means ± s.d.; data in **b** and Extended Data Fig. 1d–f, j–l were from large experimental sets sharing immunoglobulin G (IgG) controls; $n = 3$ biologically independent experiments (**b**–**f**); two-tailed t-test (**b**); one-way analysis of variance (ANOVA) with Dunnett's multiple comparisons test (**c**–**e**); image in **a** is representative of two independent experiments.

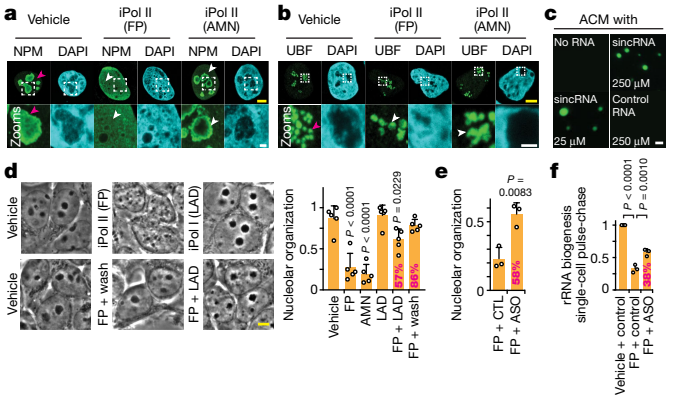


Fig. 2 | Pol II represses sincRNAs to maintain nucleolar structure and function. a, **b**, Effects of a 3-hour Pol II inhibition on NPM (**a**) and UBF (**b**) localization, as shown by immunofluorescence microscopy. Examples of normal and defective phenotypes are respectively marked by magenta and white arrowheads. DAPI, 4′,6-diamidino-2-phenylindole. **c**, Low-complexity sincRNA, not high-complexity control RNA, promoted the formation of liquid droplets in the presence of amyloid-converting motif (ACM) peptides in vitro. Shown on the images is the concentration of ACM peptides incubated with 1 µM of the indicated RNA. **d**–**f**, In cells subjected to Pol II inhibition (FP), nucleolar organization was restored by coinhibition of Pol I (LAD; **d**), removal of FP (wash; **d**), or treatment with sincRNA-repressing ASOs (**e**), which also restored rRNA biogenesis as indicated by live single-cell FU-RNA pulse-chase assays (**f**). Percentages indicating phenotypic rescue relative to FP-treated cells are shown on graphs as applicable. **a**–**f**, Experiments with HEK293T cells; data are shown as means ± s.d.; one-way ANOVA with Dunnett's multiple comparisons test (**d**, **f**); two-tailed t-test (**e**); $n = 5$ biologically independent experiments (**d**); $n = 3$ biologically independent experiments (**e**, **f**); images in **a**–**c** are representative of two independent experiments; scale bars, 5 µm (yellow) and 1 µm (white).

The Pol II activator cyclin-dependent kinase 9 (CDK9) was similarly enriched across IGSs (Extended Data Fig. 1g). pS2 and CDK9 were also enriched across the IGSs of IMR90 fibroblasts, indicating that enrichments are not limited to tumorigenic cells (Extended Data Fig. 1h, i). Unlike Pol II and CDK9, Pol I and its initiation factor, upstream binding factor 1 (UBF, also known as UBF1), localized primarily to rRNA genes, although low Pol I levels existed across IGSs (Extended Data Fig. 1j, k). Notably, Pol II was overrepresented relative to Pol I only within IGSs (Extended Data Fig. 1l). These data suggest that rDNA loci are cohabited by Pol I and Pol II.

To determine whether rRNA biogenesis is rapidly affected following Pol II perturbation, we conducted a three-hour treatment using the Pol II inhibitors α-amanitin (AMN) or flavopiridol in pulse-chase experiments. Pol II inhibition perturbed global ribosome biogenesis (Fig. 1c). Specifically, unlike Pol I inhibition by low-dose actinomycin-D (LAD), Pol II inhibition almost fully abolished pre-rRNA processing (Fig. 1d, e and Extended Data Fig. 2a–c), indicating a distinct mechanism of ribosome biogenesis arrest. Cell viability and global protein levels were unchanged following Pol II inhibition, arguing against indirect effects (Extended Data Fig. 2d, e). Furthermore, a 30-min Pol II inhibition was sufficient to strongly disrupt rRNA processing, suggesting a direct function for Pol II through its enrichment at rDNA (Extended Data Fig. 2f; Pol II inhibition hereafter was for three hours unless otherwise indicated). These data suggest that Pol II might directly support nucleolar rRNA expression through its association with IGSs.

In different cell types, we detected IGS noncoding RNAs (ncRNAs) that decreased in abundance following Pol I inhibition (Fig. 1f and Extended Data Fig. 2g, h). Strikingly, IGS ncRNAs were markedly induced and found to be de novo transcribed upon Pol II inhibition (Fig. 1f and Extended Data Fig. 2h, i). Simultaneous inhibition of Pol I abolished the induction of IGS ncRNAs by Pol II inhibition (Extended Data Fig. 2j, k). Thus, Pol II counters Pol-I-dependent synthesis of IGS ncRNAs. Strand-specific transcript analysis of IGSs identified sense intergenic ncRNAs (sincRNAs) and antisense intergenic ncRNAs (asincRNAs) that were transcribed by Pol I and Pol II, respectively (Extended Data Fig. 2l–n). The sincRNA/asincRNA ratio paralleled Pol I/Pol II enrichment across IGSs (Extended Data Fig. 2m, o). The data so far indicate that Pol II operates directly across the IGSs, where it generates asincRNAs and limits the spurious synthesis of sincRNAs by Pol I.

Pol II maintains nucleoli via sincRNA control

Given that nucleolar organization is essential for rRNA synthesis and processing, we characterized disordered proteins at the nucleolar subdomains that are essential for these functions (Extended Data Fig. 3a, b). NPM delineates the granular component of the nucleolus, the LLPS of which is required for rRNA processing[2,6,7]. Pol II inhibition abrogated the phase separation of NPM, which was quickly reorganized into ruffled bodies before undergoing complete mixing with the nucleoplasm (Fig. 2a and Extended Data Fig. 3c–e). At nucleolar fibrillar centres, UBF (which is enriched at the promoters of rRNA genes) forms small foci[6]. Pol II inhibition resulted in UBF relocation to the nucleolar periphery, where UBF formed large spheres, rings or crescent-shaped bodies exhibiting wetting behaviour (Fig. 2b and Extended Data Fig. 3f–h). Changes in NPM and UBF coincided with global nucleolar disorganization (Extended Data Fig. 3i) and matched sincRNA induction kinetics (Fig. 1f). UBF bodies generated upon Pol II inhibition exhibited greater fluorescence recovery after photobleaching

(FRAP; Extended Data Fig. 3j)[6], suggesting decreased UBF–rDNA interactions or rDNA relocation to less viscous environments. Notably, the former nucleolar space that became surrounded with UBF signals following Pol II inhibition showed positive staining with Congo red (Extended Data Fig. 3k), indicating the presence of stress-induced, solid-like nucleolar amyloid bodies[8,9]. The data suggest that Pol II inhibition partly and strongly disrupts the organization of rRNA synthesis and processing sites, respectively. Under these conditions, aberrant liquid-to-solid phase transitions occur within the remnant nucleolar space.

Nucleolar amyloid bodies usually emerge following environmental stresses such as heat shock[6,8,9]. Specifically, heat shock causes proteins with the amyloid-converting motif (ACM) to form nucleolar liquid droplets, which undergo phase transition into solid-like amyloid bodies (Extended Data Fig. 4a, b). Knockdown of different sincRNAs prevented heat-shock-induced formation of ACM-containing nucleolar liquid droplets in vivo (Extended Data Fig. 4c). In a cell-free in vitro system, incubating ACM peptides with a sincRNA segment induced liquid droplet formation (Fig. 2c and Extended Data Fig. 4d)[9]. Moreover, strand-specific RNA sequencing (ss-RNA-seq) revealed that heat shock induced sincRNA and repressed asincRNA levels at IGSs (Extended Data Fig. 4e). Thus, environmental stress represses asincRNA levels and promotes sincRNA-dependent nucleolar remodelling. The results also show that sincRNAs induce liquid droplets in vitro and promote liquid droplets and consequent solid-like amyloid bodies in vivo.

Next, we assessed whether sincRNA repression restores nucleolar organization and function in live cells subjected to Pol II inhibition. Nucleolar organization was restored after Pol II inhibitor wash-off, Pol I co-inhibition, or direct repression of sincRNA levels with antisense oligonucleotides (ASOs) (Fig. 2d, e and Extended Data Fig. 5a). ASOs also partly restored rRNA biogenesis (Fig. 2f). An overexpressed sincRNA localized to nucleoli without decreasing rRNA biogenesis (Extended Data Fig. 5b–d), indicating that nucleolar disruption may depend on specific combinations of sincRNAs or that endogenous sincRNAs have distinctive modifications or interactors. However, cell types with naturally elevated sincRNA levels exhibited more NPM-marked nucleoli (Extended Data Figs. 2g, 5e). Of note, long-term Pol II inhibition may compromise nucleoli indirectly, by limiting the ability of Pol II to synthesize the U8 small nucleolar RNA (snoRNA) or Alu RNA molecules[10–12]. However, following our short-term Pol II inhibition, nucleolar disruption coincided with sincRNA induction in the absence of changes in U8 or Alu levels (Extended Data Fig. 6a, b). Additionally, in contrast with Pol II inhibitors, pharmacological agents[2,12] disrupting nucleolar organization or global protein translation failed to induce sincRNA levels (Extended Data Fig. 6c–e). Thus, sincRNA accumulation drives nucleolar disorganization, and not vice versa. Together, these results show that Pol II constitutively represses different Pol-I-dependent sincRNAs to prevent unscheduled stress-mimicking nucleolar phase transitions, and to maintain endogenous nucleolar condensates that are essential for rRNA biogenesis.

Pol II sets an R-loop shield for Pol I

Nucleoli are naturally enriched in R-loops, which are triplex nucleic acid structures harbouring a DNA–RNA hybrid and single-stranded DNA[13]. Therefore, we postulated that baseline R-loop levels across IGSs may have beneficial effects through the modulation of Pol I–Pol II crosstalk. DNA–RNA hybrid immunofluorescence (DRIF) revealed nucleolar R-loops that were partly repressed by Pol II inhibition (Fig. 3a) or the recombinant DNA–RNA hybrid repressor RNase H1 (Extended Data Fig. 7a–c). DNA–RNA hybrid immunoprecipitation (DRIP) revealed that several IGS sites exhibited R-loop signals, which peaked at the junctions between rRNA genes and IGSs and were sensitive to RNase H1 (Fig. 3b and Extended Data Fig. 7d)[14]. Despite markedly higher transcription of rRNA genes relative to IGSs, negative GC skews may be one

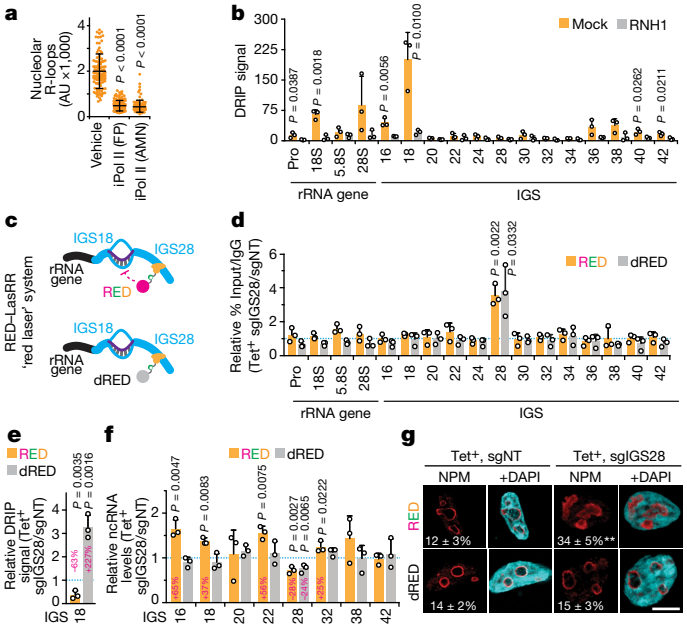


Fig. 3 | Repression of an IGS R-loop shield disrupts nucleoli. a, Pol II inhibition repressed nucleolar R-loops. **b**, DRIP analysis shows RNase H1-sensitive R-loop peaks at rDNA. **c**, The RED–LasRR system created to achieve inducible locus-associated R-loop repression. **d**, The short guide RNA for IGS28 (sgIGS28) enriched the tetracycline (Tet$^+$)-induced RED or dRED at IGS28 in anti-GFP ChIP, using IgG as control. Enrichments are normalized to a non-targeting control (sgNT). RED and dRED data were from different experiments but are shown on the same graph as a space-saving measure. **e**, Using RED or dRED together with sgIGS28 respectively decreased or increased R-loop levels at IGS18. **f**, **g**, RED sgIGS28 induced ncRNA levels (**f**) and disrupted NPM localization (**g**). The percentages of cells exhibiting ruffled NPM localization are indicated on the images (**g**). **a**–**g**, HEK293T cells; data are shown as means ± s.d.; two-tailed Mann–Whitney *U*-test, $n = 100$ cells (**a**); or two-tailed *t*-test, $n = 3$ biologically independent experiments (**b**, **d**–**f**); scale bar, 5 µm. Percentage changes relative to respective sgNT samples are indicated above or on bars (**e**, **f**).

of several different factors favouring antisense IGS R-loops (Extended Data Fig. 7e)[15]. Notably, R-loop repression by RNase H1 overexpression partly mimicked Pol II inhibition, increasing sincRNA expression at most IGS sites tested (Extended Data Fig. 7f, g). Together, these findings suggest that R-loops are important molecular mediators of sincRNA repression by Pol II.

RNase H1 overexpression remains the gold-standard method by which to interrogate R-loop function[13]. However, with this approach, RNase H1 is often not enriched at the studied loci, where the observed phenotypic changes may also be due to R-loop repression elsewhere. To specifically interrogate the function of IGS-associated R-loops, we created a tetracycline-inducible RNase H1–eGFP–dCas9 (RED) fusion protein to achieve locus-associated R-loop repression (a process that we abbreviate as 'RED–LasRR', or 'red laser'; Fig. 3c and Extended Data Fig. 7h; eGFP is enhanced green fluorescent protein). As a control, this system uses a similar chimaeric protein that comprises catalytically dead RNase H1 (denoted dRED).

Similar to the RNase H1 protein[16], RED and dRED displayed nucleolar and nucleoplasmic localization in the absence of short guide RNAs (sgRNAs) (Extended Data Fig. 7i, j). Within the IGS, constitutive chromatin looping juxtaposes the IGS27/28 sites with IGS16/18 sites[17,18]. Therefore, we investigated whether a pool of three sgRNAs targeting IGS28 (sgIGS28) can enrich RED at IGS28 and repress the strong R-loop peaks at IGS16/18. ChIP confirmed successful targeting and similar enrichment of RED and dRED at the IGS28 site upon coexpression of sgIGS28 (Fig. 3d). Targeting RED, but not dRED, to IGS28 repressed only

the strong R-loop peak at IGS18, while inducing a subset of sincRNAs across the IGSs (Fig. 3e–f and Extended Data Fig. 7k). Using RED with sgIGS38, which is spatially distal to the IGS18 site[17,18], failed to alter R-loop or ncRNA levels at IGS18 (Extended Data Fig. 7l, m). Targeting dRED to IGS28 stabilized R-loops without decreasing sincRNA levels at IGS18, suggesting that maximal function of IGS18 R-loops is already achieved endogenously (Fig. 3e, f). Of note, ncRNA levels were similarly decreased at the IGS28 site to which RED or dRED was targeted without affecting Pol II enrichments (Fig. 3f and Extended Data Fig. 7n), and the RED–LasRR system can be used to target the fusion proteins to a single-copy locus outside of rDNA (Extended Data Fig. 7o). Using the guide RNAs (gRNAs) targeting RED to IGS28, individually, failed to achieve R-loop repression at the IGSs (Extended Data Fig. 7p). This argues against the possibility that targeting of RED to non-rDNA sites via any single gRNA or the RNase H1 moiety of the fusion protein indirectly represses IGS R-loops. Although the RED/sgIGS28-dependent lowering of R-loops only partially induces sincRNAs, this still mimicked early Pol II inhibition, as shown by the perturbation of NPM architecture into indistinct, ruffled bodies (Fig. 3f, g). This highlights the disruptive impact that even small increases in sincRNA levels can exert on nucleoli. The data show that asincRNAs generated by Pol II form an antisense R-loop shield that limits the synthesis of Pol-I-dependent sincRNAs, which can abrogate nucleolar organization and function. The RED–LasRR system will support studies on the numerous roles of R-loops in genome expression and stability.

Senataxin supports the R-loop shield

We next set out to identify additional factors that may regulate nucleolar Pol II. Senataxin (SETX) is a human neurodegeneration-linked helicase[19]. SETX and its yeast orthologue Sen1 have several transcription-modulatory roles, including Pol II loading and R-loop repression[20–22]. Sen1 associates with rDNA IGSs to promote Pol I transcription termination and to silence lifespan-shortening IGS ncRNAs[16,23,24]. We found that SETX was enriched across human IGSs, especially at IGS28, and exhibited nucleolar localization (Extended Data Fig. 8a, b). The IGS28 SETX peak overlapped one Pol II peak and the intergenic promoter marks H3K27ac, H3K9ac and H3K4me3 in ENCODE ChIP–seq data (Extended Data Fig. 8c). Sequential ChIP revealed that SETX was preferentially coenriched with Pol II compared with Pol I at IGS28 (Fig. 4a). Thus, SETX is coenriched with Pol II at IGSs, especially at a putative intergenic promoter at IGS28.

Notably, SETX knockout decreased the intergenic enrichment of Pol II and its R-loops (Fig. 4b and Extended Data Fig. 8d, e). This change was accompanied by increased intergenic Pol I enrichment (Extended Data Fig. 8f), elevated sincRNA synthesis (Fig. 4c and Extended Data Fig. 8g), and decreased Pol I localization at rRNA genes (Extended Data Fig. 8h). Unlike SETX knockout, the forced release of Pol I from rRNA gene promoters through knockdown of transcription initiation factor 1A (TIF1A) decreased pre-rRNA levels without inducing sincRNA levels (Extended Data Fig. 8i, j). This suggests that SETX loss prevents Pol II from shielding the IGSs from de novo Pol I loading. In addition, northern blotting did not show increases in pre-rRNA length upon Pol II or SETX disruption, arguing against rRNA gene read-through as the basis for increased IGS transcription by Pol I (Extended Data Fig. 8k). Thus, IGS R-loops act more as a shield that prevents Pol I recruitment, rather than a barrier that limits read-through transcription. Increases in sincRNA levels in SETX-knockout cells were associated with nucleolar disorganization and pre-rRNA processing defects, which were partly countered by sincRNA knockdown (Fig. 4d, e and Extended Data Fig. 8l–n). That SETX loss partly mimicked Pol II inhibition probably reflects the partial coenrichment of SETX and Pol II at IGSs. Additionally, SETX knockout did not lower IGS epigenetic silencing marks (Extended Data Fig. 8o), suggesting that SETX loss does not promote sincRNA levels by abrogating epigenetic silencing. In fact, SETX knockout slightly increased

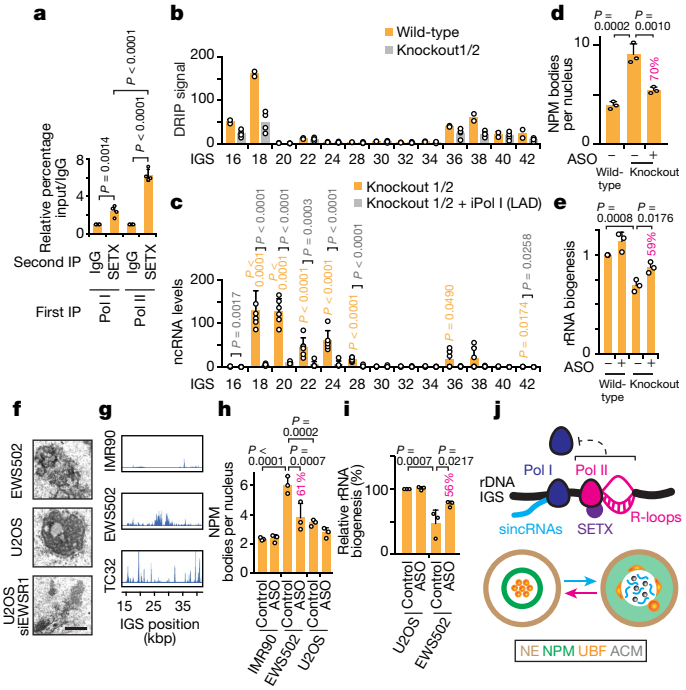


Fig. 4 | Nucleolar Pol II reinforcement by SETX and nucleolus-disrupting sincRNAs in cancer. a, Sequential immunoprecipitations (IPs) revealed preferential coenrichment of SETX with Pol II at IGSs. Signals for Pol I/SETX and Pol II/SETX immunoprecipitations are normalized to signals from Pol I/IgG and Pol II/IgG, respectively. **b**, **c**, SETX knockout in two clones decreased R-loops (**b**) and induced IGS ncRNAs (**c**). **d**, **e**, Single-cell analysis of SETX-knockout cells showed that ASO-mediated repression of Pol-I-dependent sincRNAs partly rescues nucleolar organization (**d**) and rRNA biogenesis (**e**). Percentages indicating the magnitude of ASO-mediated phenotypic rescue are shown above graph bars where applicable. **f**, The patient-derived Ewing sarcoma cell line EWS502 and U2OS osteosarcoma cells with siRNA-mediated depletion of EWS breakpoint region 1 (EWSR1) showed disrupted nucleoli by electron microscopy. **g**, RNA-seq data indicate increased ncRNA levels at the IGSs of EWS502 and TC32 cells, as compared with IMR90 control cells. kbp, kilobasepairs. **h**, **i**, Single-cell analysis showed that sincRNA knockdown partly restores nucleolar organization (**h**) and rRNA biogenesis (**i**) in EWS502 cells. **j**, Model showing how a Pol-II-dependent R-loop shield limits Pol-I-dependent sincRNAs, which compromise nucleolar organization and function. ACM, amyloid-converting motif; NE, nuclear envelope. **a**–**i**, Cells were HEK293T (**a**–**e**) or as indicated (**f**–**i**); data are shown as means ± s.d.; one-way ANOVA with Dunnett's multiple comparisons test (**a**, **d**, **e**, **h**, **i**) and one-way ANOVA with Tukey's multiple comparisons test (**c**); $n = 4$ biologically independent experiments (**a**), $n = 2$ biologically independent experiments (**b**, duplicates for each of wild-type, knockout 1 and knockout 2), $n = 6$ biologically independent experiments (**c**, triplicates for each knockout), $n = 3$ biologically independent experiments (**d**, **e**, **h**, **i**); images in **f**, **g** are representatives of two independent experiments; scale bar, 1 µm.

silencing marks, possibly reflecting epigenetic compensation constraining the magnitude of sincRNA induction. The data indicate that SETX is coenriched with IGS Pol II and supports it in repressing a subset of Pol-I-dependent sincRNAs that can disrupt nucleolar organization and function. SETX may achieve this effect by promoting the efficient loading and release of Pol II at an IGS28 intergenic promoter.

sincRNAs can disrupt nucleoli in cancer

We then aimed to identify a setting in which naturally elevated sincRNA levels may compromise nucleolar structure and function. Nucleolar organization, which is intimately related to cellular growth and viability, may be an adjunct in the diagnosis and treatment of some cancers[25]. In fact, nucleolar disruption upon Pol II dysregulation is similar to the

constitutive disorganization of nucleoli in human Ewing sarcoma (EWS) tumours, related patient-derived EWS502 or TC32 cells, and U2OS osteosarcoma cells with depletion of EWS breakpoint region 1 (EWSR1) (Fig. 4f and Extended Data Fig. 9a, b)[26]. To determine whether alterations in sincRNA levels could underlie this phenotype, we reanalysed RNA-seq and DRIP–seq data from EWS and healthy IMR90 control cells to include rDNA[17,26]. EWS cells exhibited increased ncRNA and R-loop levels across IGSs (Fig. 4g and Extended Data Fig. 9c–e). Strikingly, in EWS cells, nucleolar disorganization and an rRNA biogenesis defect were countered by sincRNA knockdown (Fig. 4h, i and Extended Data Fig. 9f). These findings suggest that natural increases in sincRNA levels can explain aberrant nucleolar morphologies that are commonly observed in cancer[25]. R-loop increases in this setting may reflect selection for cells that have compensated for the increased sincRNA levels.

Discussion

Our findings indicate that, in mammalian cells, antisense transcription by nucleolar Pol II generates an R-loop shield at rDNA IGSs to block Pol-I-dependent sense intergenic transcripts, which can compromise nucleolar condensates underlying rRNA expression (Fig. 4j and Extended Data Fig. 10). Processes that restrain R-loops at human IGSs probably exist, as unrestricted IGS R-loops destabilize yeast rDNA[4,16,27]. However, our findings differ from those in yeast, where IGS transcription does not regulate rRNA[28,29] and Sen1 limits deleterious IGS ncRNAs by enforcing epigenetic silencing and transcript turnover[21,23]. At the IGSs of human cells under stress[8], protective sense RNAs are likely to be induced through local repression of antisense RNA and R-loops. Nucleolar Pol II at IGSs may also mediate crosstalk with cellular differentiation, which is partly driven by promoter-associated transcripts that are dependent on Pol I or Pol II[30–33]. Future work should explore the potential use of sincRNAs and nucleolar disorganization as cancer biomarkers, and whether tumours exhibiting such features are hypersensitive to Pol-II-inhibiting drugs[34]. Overall, we identify nucleolar Pol II as a new master regulator of ribosome biogenesis, with broad implications for health and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2497-0.

1. Boisvert, F. M., van Koningsbruggen, S., Navascués, J. & Lamond, A. I. The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.* **8**, 574–585 (2007).
2. Feric, M. et al. Coexisting liquid phases underlie nucleolar subcompartments. *Cell* **165**, 1686–1697 (2016).
3. Mekhail, K. & Moazed, D. The nuclear envelope in genome organization, expression and stability. *Nat. Rev. Mol. Cell Biol.* **11**, 317–328 (2010).
4. Kobayashi, T. & Ganley, A. R. Recombination regulation by transcription-induced cohesin dissociation in rDNA repeats. *Science* **309**, 1581–1584 (2005).
5. Kaeberlein, M., McVey, M. & Guarente, L. The SIR2/3/4 complex and SIR2 alone promote longevity in *Saccharomyces cerevisiae* by two different mechanisms. *Genes Dev.* **13**, 2570–2580 (1999).
6. Mekhail, K. et al. Regulation of ubiquitin ligase dynamics by the nucleolus. *J. Cell Biol.* **170**, 733–744 (2005).
7. Mitrea, D. M. et al. Self-interaction of NPM1 modulates multiple mechanisms of liquid-liquid phase separation. *Nat. Commun.* **9**, 842 (2018).
8. Audas, T. E. et al. Adaptation to stressors by systemic protein amyloidogenesis. *Dev. Cell* **39**, 155–168 (2016).
9. Wang, M. *et al*. Stress-induced low complexity RNA activates physiological amyloidogenesis. *Cell Rep.* **24**, 1713–1721 (2018).
10. Caudron-Herger, M. et al. Alu element-containing RNAs maintain nucleolar structure and function. *EMBO J.* **34**, 2758–2774 (2015).
11. Burger, K. et al. Cyclin-dependent kinase 9 links RNA polymerase II transcription to processing of ribosomal RNA. *J. Biol. Chem.* **288**, 21173–21183 (2013).
12. Burger, K. et al. Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. *J. Biol. Chem.* **285**, 12416–12425 (2010).
13. Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.* **16**, 583–597 (2015).
14. Bhatia, V. et al. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**, 362–365 (2014).
15. Sanz, L. A. et al. Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell* **63**, 167–178 (2016).
16. Abraham, K. J. et al. Intersection of calorie restriction and magnesium in the suppression of genome-destabilizing RNA-DNA hybrids. *Nucleic Acids Res.* **44**, 8870–8884 (2016).
17. Zentner, G. E., Saiakhova, A., Manaenkov, P., Adams, M. D. & Scacheri, P. C. Integrative genomic analysis of human ribosomal DNA. *Nucleic Acids Res.* **39**, 4949–4960 (2011).
18. Shiue, C. N., Nematollahi-Mahani, A. & Wright, A. P. Myc-induced anchorage of the rDNA IGS region to nucleolar matrix modulates growth-stimulated changes in higher-order rDNA architecture. *Nucleic Acids Res.* **42**, 5505–5517 (2014).
19. Groh, M., Albulescu, L. O., Cristini, A. & Gromak, N. Senataxin: genome guardian at the interface of transcription and neurodegeneration. *J. Mol. Biol.* **429**, 3181–3195 (2017).
20. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell* **42**, 794–805 (2011).
21. Steinmetz, E. J. et al. Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol. Cell* **24**, 735–746 (2006).
22. Suraweera, A. et al. Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. *Hum. Mol. Genet.* **18**, 3384–3396 (2009).
23. Vasiljeva, L., Kim, M., Terzi, N., Soares, L. M. & Buratowski, S. Transcription termination and RNA degradation contribute to silencing of RNA polymerase II transcription within heterochromatin. *Mol. Cell* **29**, 313–323 (2008).
24. Mischo, H. E. et al. Yeast Sen1 helicase protects the genome from transcription-associated instability. *Mol. Cell* **41**, 21–32 (2011).
25. Stamatopoulou, V., Parisot, P., De Vleeschouwer, C. & Lafontaine, D. L. J. Use of the iNo score to discriminate normal from altered nucleolar morphology, with applications in basic cell biology and potential in human disease diagnostics. *Nat. Protoc.* **13**, 2387–2406 (2018).
26. Gorthi, A. et al. EWS-FLI1 increases transcription to cause R-loops and block BRCA1 repair in Ewing sarcoma. *Nature* **555**, 387–391 (2018); correction **559**, E11 (2018).
27. El Hage, A., French, S. L., Beyer, A. L. & Tollervey, D. Loss of topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev.* **24**, 1546–1558 (2010).
28. Kawauchi, J., Mischo, H., Braglia, P., Rondon, A. & Proudfoot, N. J. Budding yeast RNA polymerases I and II employ parallel mechanisms of transcriptional termination. *Genes Dev.* **22**, 1082–1092 (2008).
29. Salvi, J. S. et al. Roles for Pbp1 and caloric restriction in genome and lifespan maintenance via suppression of RNA-DNA hybrids. *Dev. Cell* **30**, 177–191 (2014).
30. Bierhoff, H. et al. Quiescence-induced LncRNAs trigger H4K20 trimethylation and transcriptional silencing. *Mol. Cell* **54**, 675–682 (2014).
31. Savić, N. et al. lncRNA maturation to initiate heterochromatin formation in the nucleolus is required for exit from pluripotency in ESCs. *Cell Stem Cell* **15**, 720–734 (2014).
32. Mayer, C., Schmitz, K. M., Li, J., Grummt, I. & Santoro, R. Intergenic transcripts regulate the epigenetic state of rRNA genes. *Mol. Cell* **22**, 351–361 (2006).
33. Zhao, Z., Sentürk, N., Song, C. & Grummt, I. lncRNA PAPAS tethered to the rDNA enhancer recruits hypophosphorylated CHD4/NuRD to repress rRNA synthesis at elevated temperatures. *Genes Dev.* **32**, 836–848 (2018).
34. Zhang, H. et al. Targeting CDK9 reactivates epigenetically silenced genes in cancer. *Cell* **175**, 1244–1258 (2018).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment, except for the quantification of microscopy images.

Cell culture and general materials

Human HEK293T, HeLa, HAP1 and osteosarcoma (U2OS) cells were cultured in Dulbecco's modified Eagle medium (DMEM, Wisent Bioproducts) with 10% fetal bovine serum (FBS, Wisent). HEK293T T-REx cells (ThermoFisher Scientific) were cultured in DMEM supplemented with 10% tetracycline-free FBS and 1% penicillin/streptomycin. EWS502 and IMR90 cells were cultured in Roswell Park Memorial Institute (RPMI) medium supplemented with 10% FBS. All cell lines were cultured in the presence of 1% (*v*/*v*) penicillin/streptomycin (Wisent) at 37 °C in a humidified atmosphere with 5% CO_2. Transfection of cultured cells was achieved using Lipofectamine3000 (Invitrogen, catalogue number L3000008), Lipofectamine RNAiMAX (Invitrogen, catalogue number 13778150) and Polyjet DNA transfection reagent (SignaGen Laboratories, SL100688). For transfections with plasmids encoding GFP–UBF1 or RNaseH1, 70% confluent cells were transfected with 1–3 μg of plasmid per well of a six-well plate; pcDNA3 served as control for RNaseH1 overexpression. For Pol II inhibition, cells were treated either with the reversible inhibitor flavopiridol (2 μM, inhibits Pol II pS2; Santa Cruz catalogue number sc-202157) or with the irreversible inhibitor α-amanitin (AMN, 50 μg ml^{-1}, inhibits translocation; Abcam catalogue number ab144512). Other drug treatments were LAD (50 ng ml^{-1}), MG132 (10 μM), doxorubicin (Dox, 300 nM), camptothecin (CPT, 10 μM), cycloheximide (CHX, 100 μM) or 1,6-hexanediol (HEX, 0.1% *v*/*v*). Antibodies, primers, guide RNAs and northern probes are listed in the Supplementary Information (Supplementary Tables 1–4). For Ewing sarcoma analyses, the Ewing sarcoma cell line TC32 was procured from the Children's Oncology group (https://childrensoncologygroup.org/) and EWS502 was a kind gift from S. Lessnick (Nationwide Children's Hospital, OH). Both cell lines were grown in RPMI (Corning). The control cell lines IMR90 (a primary fibroblast cell line) and U2OS (a human osteosarcoma cell line) were purchased from the American Type Culture Collection (ATCC) and grown in DMEM (Corning). Media were supplemented with 10% heat-inactivated FBS (Atlanta Biologicals). Cells were maintained at 37 °C in a humidified atmosphere with 5% CO_2, confirmed using short tandem repeat (STR) profiling and tested for mycoplasma contamination. All siRNA transfections were conducted using Lipofectamine RNAimax (ThermoFisher Scientific) according to the manufacturer's protocols.

Chromatin immunoprecipitation

Cells were grown to 80% confluence in 15 cm plates and crosslinked by adding 1% (*v*/*v*) formaldehyde at room temperature for 10 min. The reaction was quenched with 125 mM glycine for 5 min at room temperature. Cells were washed twice with cold phosphate-buffered saline (PBS), lysed with 10 ml lysis buffer (5 mM PIPES, 85 mM KCl, 0.5% (*v*/*v*) NP-40, complete protease-inhibitor cocktail (Roche)), scraped into tubes, and incubated for 10 min on ice. Cells were then pelleted at 1,000 r.p.m. for 10 min at 4 °C and resuspended in 500 μl of nuclear lysis buffer (50 mM Tris-HCl, 10 mM EDTA, 10% (*w*/*v*) SDS, complete protease-inhibitor cocktail) and incubated on ice for 10 min. Lysates were sonicated eight times for 20 s each at 40% amplitude at 4 °C with intermittent incubations on ice for 2 min. Centrifugation at 12,500*g* for 10 min at 4 °C clarified lysates. We set aside 10 μl of sheared chromatin for each sample as input. We diluted 50 μl of chromatin at a 1/10 ratio in immunoprecipitation dilution buffer (16.7 mM Tris-HCl pH 8.0, 0.01% (*w*/*v*) SDS, 167 mM NaCl, 1.2 mM EDTA, 1.1% (*v*/*v*) Triton-X100, complete protease inhibitor) and incubated with 5 μg of antibody on a rotator overnight at 4 °C. Samples were then incubated at constant rotation with 25 μl of prewashed Dynabeads protein G (Life Technology, catalogue number 10004D) for 2 h at 4 °C. Beads were washed once with a low-salt wash buffer (20 mM Tris-HCl, 0.1% (*w*/*v*) SDS, 1% (*v*/*v*) Triton X-100, 2 mM EDTA, 150 mM NaCl), once with high-salt wash buffer (20 mM Tris-HCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl), once with LiCl wash buffer (10 mM Tris-HCl, 1% (*v*/*v*) NP-40, 1% (*w*/*v*) sodium deoxycholate, 1 mM EDTA, 250 mM LiCl), and twice with TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) before two rounds of incubation with 100 μl of elution buffer (1% SDS, 100 mM $NaHCO_3$) for 15 min at room temperature. The eluates were incubated with 8 μl of 5 M NaCl on a rotator at 65 °C overnight. We added 3 μl of 10 mg ml^{-1} RNase A (ThermoFisher Scientific, catalogue number EN0531) and incubated samples first at room temperature for 30 min, and then with 4 μl of 0.5 M EDTA, 8 μl of 1 M Tris-HCl and 1 μl proteinase K (Roche, catalogue number 03115887001) at 45 °C for 2 h. DNA was purified using gel/PCR DNA-fragment extraction (Geneaid, catalogue number DF300) and diluted with 150 μl of TE buffer. Primers are listed in the reagents table included with the Supplementary Information. Following ChIP–qPCR analysis, ChIP enrichments, accounting for typical background fluctuations across repetitive DNA loci, were calculated as (Percentage of input/IgG) = (Percentage of input for protein immunoprecipitation)/(Percentage of input for mock IgG immunoprecipitation). The mean IgG background is also shown on ChIP graphs (Fig. 1b and Extended Data Figs. 1d–k, 8a).

Sequential chromatin immunoprecipitation

Similar to regular ChIP, for sequential chromatin immunoprecipitation (ChIP–reChIP), cells were grown to 80% confluence, crosslinked and lysed. For the first round of immunoprecipitation, samples were diluted 1/10 in immunoprecipitation dilution buffer (100 μl chromatin plus 900 μl immunoprecipitation dilution buffer) and incubated with 5 μg of anti-Pol-I (anti-RPA135 subunit) or anti-Pol-II (anti-C-terminal domain (CTD)) on a rotator overnight at 4 °C. Samples were then incubated at constant rotation with 25 μl of pre-washed Dynabeads for 2 h at 4 °C. Similar to regular ChIP (see above), beads were washed once with low-salt wash buffer, once with high-salt wash buffer, once with LiCl wash buffer, and twice with TE buffer before one 30-min incubation with 50 μl elution buffer containing 10 mM DTT. Eluates from each of the first immunoprecipitation tubes corresponding to the same antibody were combined, diluted 20-fold in cold immunoprecipitation dilution buffer and incubated overnight at 4 °C with 5 μg of anti-senataxin (anti-SETX) antibody. Once again, beads were incubated at constant rotation with 25 μl of pre-washed Dynabeads for 2 h at 4 °C, washed once with low-salt wash buffer, once with high-salt wash buffer, once with LiCl wash buffer, and twice with TE buffer before two rounds of incubation with 100 μl of elution buffer for 15 min at room temperature, and overnight incubation at 65 °C with 8 μl of 5 M NaCl. Similar to regular ChIP, samples were treated with RNase A/proteinase K (Roche, catalogue number 03115887001) and purified; qPCR was then performed.

Quantitative PCR

Quantitative real-time PCR was performed using a Bio-Rad CFX Connect Real-Time. Ten microlitres of qPCR reactions each containing SensiFAST SYBR No-ROX kit (FroggaBio, catalogue number BIO-98050), 200 nM of each of the forward and reverse primers, and 1 μl of diluted complementary DNA, diluted input, diluted immunoprecipitation ChIP or diluted DRIP DNA depending on the experiment. PCR comprised one cycle of 95 °C for 5 min and 60 °C for 30 s, followed by 39 cycles of 95 °C for 5 s and 60 °C for 30 s, and a final melt curve of 65 °C to 95 °C in 0.5 °C steps at 5 s per step.

RNA extraction

Cells grown to 70–80% confluence were washed with RNase-free PBS before RNA isolation using a Qiagen RNeasy mini Kit (catalogue number 74104).

Reverse transcription

For regular reverse transcription, 1 µg of total RNA was treated with 1 µl of 10× DNase-I reaction buffer and 1 µl of DNase I Amp grade (1 U μl^{-1}; ThermoFisher, catalogue number 18068015), and then incubated for 15 min at room temperature. The reaction was quenched with 1 µl of 25 mM EDTA and incubated for 10 min at 65 °C. We carried out 10 µl reverse-transcription reactions using 10 mM deoxynucleoside trisphophate (dNTPs), 50 µM random nonamers (Sigma, catalogue number R7647), 500 ng total RNA, 5× first-strand buffer, 100 mM dithiothreitol (DTT), 40 U μl^{-11} RNaseOUT (Invitrogen, catalogue number 10777019) and 200 U μl^{-1} M-MLV reverse transcriptase (Invitrogen, catalogue number 28025013) at 25 °C for 10 min, 37 °C for 60 min and 70 °C for 15 min. For pre-rRNA pulse chase, an additional step comprising 5 min at 85 °C was added to release the RNA from beads. The reverse-transcription reaction was diluted 1:5, and 4 µl were used in qPCR amplification. For strand-specific (ss)RT–qPCR, 30 µg of total RNA was treated with DNase I (10 U DNase I in a 100 µl reaction) for 30 min at 37 °C. The reaction was stopped by adding 2 µl of 250 mM EDTA, pH 8.0, and incubating at 75 °C for 10 min. RNA was precipitated with 25 µl RNA precipitation solution (0.8 M trisodium citrate, 1.2 M NaCl) and 50 µl isopropanol. Samples were incubated for 10 min at room temperature and 20 min at −20 °C, and then centrifuged at 7,500*g* for 20 min at 4 °C. Supernatants were aspirated and pellets were air-dried for 10 min. Pellets were resuspended in 30 µl deionized, diethylpyrocarbonate-treated (ddDEPC) H_2O and incubated at 65 °C for 5 min. Concentrations of purified RNA were measured using NanoDrop.

We designed strand-specific primers to allow the detection of sense and antisense transcripts at the same locus as described previously[35]. Briefly, a primer of roughly 18 bp was designed to recognize the strand of interest (for example, a reverse primer to detect sense transcripts, a forward primer to detect antisense transcripts). A nonsense sequence (CGAGGATCATGGTGGCGAATAA) was added to tag the 5′-end of each strand-specific IGS primer. As a control within each reverse-transcription reaction, we generated a reverse primer to detect 7SK sense transcripts (we added a T7 sequence to the 5′ end of this primer to distinguish it from IGS primers). Separate reverse-transcription reactions were carried out for each transcript of interest. Each 10 µl reverse-transcription reaction contained 200 ng purified RNA, 5 µM strand-specific tagged primer (comprising roughly 18 bp specific to the transcript of interest, with the nonsense sequence CGAGGATCATGGTGGCGAATAA added to the 5′-end), 5 µM control sense primer (for example, 7SK), 1 mM dNTPs, 1× first-strand buffer, 10 mM DTT, 40 U RNaseOUT and 200 U of M-MLV reverse transcriptase. False-prime reactions were also carried out for each RNA sample and were conducted by replacing the transcript-of-interest primers with DEPC ddH_2O. Reactions were incubated at 25 °C for 10 min, 37 °C for 60 min, and 70 °C for 15 min. Resulting cDNA was diluted 1 in 10. Each cDNA sample represents one strand-specific transcript of interest and 7SK sense transcripts as a control. Each cDNA sample was amplified using primers directed at the strand-specific transcript of interest (using ss_Tag and hIGS_forward primers for sense transcripts or ss_Tag and hIGS_reverse primers for antisense transcripts), as well as 7SK (using T7 and 7SK forward primers). False-primed cDNA was amplified using all primer sets. qPCR reactions were performed at 95 °C for 5 min and 60 °C for 30 s, followed by 39 cycles of 95 °C for 5 s and 60 °C for 30 s. Results were analysed using the following formula: $\Delta\Delta Ct = 2\text{^-}(\Delta CtMutant - \Delta CtWT)$, where $\Delta Ct = Ct_{(transcript\ of\ interest)} - Ct_{(control)}$, and Ct is the cycle threshold. Values were normalized to those of false-prime reactions.

Population-level pre-rRNA pulse-chase

Click-iT Nascent RNA capture (Invitrogen, catalogue number C10365) was used. Cells were seeded in six-well dishes at 500,000 cells per well and allowed to grow to 40–50% confluence. Twenty-four hours later, cells were incubated with 0.15 mM ethyl uridine (EU) for 1 h, then with EU-free media for 2.5 h. Total RNA was extracted using Qiagen RNeasy kit (Qiagen, catalogue number 74104), and 1 µg of extracted RNA was incubated with 25 µl Click-iT EU buffer, 4 µl $CuSO_4$, 1.25 µl biotin azide, 1.25 µL Click-iT reaction buffer additive 1 for 3 min before addition of 1.25 µl Click-iT reaction buffer additive 2 and incubation for 30 min. The reaction mix was then incubated with 1 µl of UltraPure Glycogen (Roche, catalogue number 10901393001), 50 µl of 7.5 M ammonium acetate, and 700 µl of chilled 100% ethanol at −80 °C overnight. RNA was then pelleted using centrifugation at 13,000*g* for 20 min at 4 °C and two rounds of washes with 700 µl of 75% ethanol. We then treated 1 µg of the RNA with 31 µl Click-iT RNA binding buffer and 2 µl RNaseOUT before incubation for 5 min at 68–70 °C. The heated RNA-binding reaction mix was incubated with 12 µl of washed bead suspension at room temperature for 30 min. The beads were washed five times with Click-iT reaction wash buffer 1 and five times with Click-iT reaction wash buffer 2. The beads were then resuspended with 12 µl of Click-iT reaction wash buffer 2 and incubated at 68–70 °C for 5 min before proceeding with reverse transcription and qPCR. Processing was measured by qPCR assessment of the levels of unprocessed pre-rRNA containing the 5′ external transcribed spacer (ETS) compared with the total levels of mature rRNA.

Single-cell rRNA biogenesis assay

On the day before the assay, cells from different experimental conditions were harvested and seeded onto poly-L-lysine (PLL)-coated coverslips in 24-well plates. On the day of the assay, live cells were pulse-labelled with 1 mM 5-fluorouracil (5-FU; Sigma, catalogue number F5130) for 15 min, gently washed with unlabelled media, and chased for 30 min. Cells were then fixed and immunostained as described in the Methods section 'Endogenous protein immunofluorescence'. Double immunofluorescence labelling of nucleolar fibrillar centres or 5-FU-labelled RNA was performed using an anti-ATXN2 or anti-BrdU antibody, respectively. Random single cells were imaged captured at 100× using a Nikon C2+ confocal microscope coupled to NIS-Elements AR software (Nikon). Images were equally and evenly contrasted and ribosome biogenesis was measured as the ratio of ATXN2-marked nucleolar fibrillar centres with surrounding rRNA rings over the total number of nuclear ATXN2 foci.

Nuclear run-on

Click-iT Nascent RNA capture (Invitrogen, catalogue number C10365) was used for nuclear run-on (NRO). The setup was similar to that in the 'Population-level pre-RNA pulse-chase' section above, except that the total RNA was extracted after a 1 h incubation with 0.15 mM EU. Similar to pulse-chase labelling, the extracted RNA was biotinylated, precipitated, washed using Dynabeads, and reverse transcribed; qPCR was performed to measure the synthesis of nascent sincRNAs.

DNA–RNA hybrid immunoprecipitation

For DNA–RNA hybrid immunoprecipitation (DRIP) experiments, cells were first seeded in 60 mm plates at 2.5×10^6 cells per ml and allowed to grow to 70% confluence. Cells were then washed twice with ice-cold PBS, scraped, and centrifuged at 253*g* for 5 min. Cell pellets were resuspended in 1.6 ml TE buffer and incubated with 41.5 µl of 20% SDS and 5 µl of proteinase K overnight at 37 °C. Then, 1.6 ml of phenol-chloroform was added to cells before centrifugation at 466*g* for 5 min at room temperature. The aqueous layer was transferred and the addition of phenol-chloroform was repeated. The DNA was then precipitated by adding a 1/10 volume of 3 M NaOAc, pH 5.2, and 2.4 volumes of 100% ethanol to the aqueous layer. The DNA fibre was washed five times with 70% ethanol, resuspended in TE buffer and incubated with 3.5 µl spermidine (Bioshop, catalogue number SPR070), 35 µl buffer 2.1 (NEB), 5 µl *Hin*dIII (NEB, R01045), 10 µl *Eco*RI (Thermo Fisher, ERO271), 10 µL *Bsr*GI (NEB, R05755), 5 µl *Xba*I (NEB, R01455) and 2 µl *Ssp*I

(NEB, R0132). We then added 40 µl of 3 M NaOAc, pH 5.2, and one volume of phenol-chloroform to the digested DNA, which was then centrifuged at maximum speed for 5 min. The aqueous layer was transferred, and addition of phenol-chloroform was repeated.

To precipitate the DNA, 2.4 volumes of cold 100% ethanol were added to the aqueous layer, incubated at −20 °C for 15 min, and centrifuged at maximum speed for 30 min at 4 °C. The DNA pellet was washed with 70% ethanol and spun at maximum speed for 5 min at 4 °C. The dry pellet was resuspended in 50 µl TE buffer, and 4.4 µg of the DNA was incubated with 350 µl TE buffer, 50 µl 10× binding buffer (100 mM $NaPO_4$ pH 7.0, 1.4 M NaCl, 0.5% (*v*/*v*) Triton X-100) and 10 µg of either mouse IgG or S9.6 antibody at 4 °C overnight. Immunoprecipitation samples were incubated with previously washed Dynabeads for 2 h at 4 °C. Samples were then washed three times with 1× binding buffer and eluted off the beads by incubation with DRIP elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 0.5% (*w*/*v*) SDS) and proteinase K for 45 min at 55 °C. The DNA was then purified using gel/PCR DNA fragment extraction (Geneaid, catalogue number DF300) and qPCR of purified DNA was performed. The specificity of the S9.6 antibody for RNA–DNA hybrids was confirmed by in vitro treatment with RNase H1 in all experiments. We also screen all antibodies for specificity by ensuring that signals do not exhibit any statistically significant changes following treatment with RNase III (NEB, catalogue number M0245S). Following ChIP–qPCR analysis, background IgG mock signal was subtracted from S9.6 immunoprecipitation signal to generate a DRIP signal, which was then plotted as a raw DRIP signal or as a relative DRIP signal when normalized to a given site or condition.

Fluorescence recovery after photobleaching

HEK293T cells were transfected with a GFP–UBF1 plasmid 24 h before cell passaging to 2 cm glass-bottomed live microscopy dishes. Next day, the roughly 75% confluent cells were treated with either flavopiridol to a final concentration of 2 µM or dimethylsulfoxide (DMSO) as a control. Cells were incubated for 3.5 h and subjected to fluorescence recovery after photobleaching (FRAP) microscopy. Confocal microscopy was executed using a ×100 oil-immersion lens (numerical aperture 1.47) on a Leica DMi8 motorized inverted microscope (Leica Microsystems) coupled to a VT-iSIM multipoint scanner (VisiTech International) and detected with a Flash 4.0 v3 sCMOS camera (Hamamatsu). FRAP was performed using the iLas FRAP system (Gattaca Systems). Design of the acquisition journals and system integration were by Quorum Technologies. Images were acquired with a 488-nm excitation wavelength laser at 15% intensity. Cells were initially imaged 20 times, and the point of interest was subsequently bleached with a 405-nm laser for 36 ms at a laser intensity of 26%. Cells were then imaged repeatedly for approximately 1 min post-bleach to capture recovery. Signal intensity was measured using MetaMorph analysis software. For analysis, the intensity of the region of interest was normalized to a nucleoplasmic background region at every time point. These background-adjusted values were then normalized to the intensity value from the first time point. The bleach time points (6–6.3 s) display saturated fluorescence as the bleached region of interest and were therefore not included in any quantification. Post-bleach control focus intensity values of greater than 1 are a result of bleach-induced decreases in nucleoplasmic background.

Creation and use of RED–LasRR system

Full-length human RNase H1 was fused to eGFP and the deactivated *Streptococcus pyogenes* Cas9 (with D10A and H840A mutations). The 5,844-nucleotide RNaseH1–SV40NLS–eGFP–SV40NLS–dCas9 gene was synthesized and cloned into the pcDNA4/TO plasmid (Invitrogen) using *Not*I and *Xba*I restriction sites (here, NLS is a nuclear localization sequence, and SV40 is simian virus 40). To ensure protein flexibility, a $(GGGS)_4$ linker was inserted between RNase H1 and the first SV40NLS, and another between eGFP and the second SV40NLS. GGS linkers were also inserted between the first SV40NLS and eGFP, and between the second SV40NLS and dCas9. dRNaseH1-eGFP-dCas9 was generated by introducing the point mutation D210N to RNase H1 using the Q5 site-directed mutagenesis kit (NEB, catalogue number E055450) according to the manufacturer's protocol, with a modification of a 15-min instead of 5-min incubation with the KLD enzyme mix at room temperature. The oligonucleotide sequences for PCR amplification were 5′-gttctgtatacaaacagtatgttt -3′ and 5′-cagtttattgatgttttgagtctt -3′. The resulting RNaseH1–SV40NLS–eGFP–SV40NLS–dCas9 (RED) or its RNase H1-dead version (dRED) was then integrated into the T-REx (ThermoFisher Scientific) tetracycline-controlled expression system. Inducible expression of the fusion proteins is thus based on the binding of tetracycline to the Tet repressor, thereby derepressing the promoter controlling the expression of the RED and dRED fusion protein. To achieve locus-specific RED and dRED–LasRR enrichment, cells were allowed to reach 70% confluence over a period of roughly 24 h. For the inducible condition, cells were incubated with medium containing tetracycline (1 µg ml^{-1}); for the uninduced condition, cells were incubated with tetracycline-free medium. All cells were transfected with 3 µg of RNH1–eGFP–dCas9 and dRED–LasRR plasmid per 60-mm plate by using Lipofectamine3000 (ThermoFisher) as per the manufacturer's instructions. Induced cells were then cotransfected using RNAiMAX (ThermoFisher) as per the manufacturer's instructions with either 4.5 µl of 10 pmol $µl^{-1}$ nontargeting sgRNA, 1.5 µl of 10 pmol $µl^{-1}$ of each of three sgRNAs for IGS18, 1.5 µl of 10 pmol $µl^{-1}$ for each of three sgRNAs for IGS28, 1.5 µl of 10 pmol $µl^{-1}$ for each of three sgRNAs for IGS38, or 1.5 µl of 10 pmol $µl^{-1}$ for each of three sgRNAs for the *β-actin 5′* pause element. The cells were incubated for 36 h before further experiments were performed.

CRISPR-mediated genome editing

For CRISPR-mediated gene knockout of SETX, CRISPR/Cas9 plasmids (pCMV–Cas9–GFP) were purchased from Sigma-Aldrich to express with the scrambled guide RNA and guide RNA for SETX (first intron). The transfections of the plasmids into the Flp-In 293 T-REx cell lines were performed with FuGENE transfection reagent (Roche, catalogue number E269A). We transfected 2 µg of the plasmid into HEK293T cells; one day after transfection, we sorted cells by BD FACSAriaTM flow cytometry (Donnelly Centre, Univ. Toronto), and plated single GFP-positive cells into 96-well plates. To confirm SETX knockout, the expression levels of SETX in each clone were detected by qPCR.

Northern blotting

RNA was prepared as described in the 'RNA extraction' and 'Reverse transcription' sections above. We then electrophorized 3.5 µg of RNA, and digoxygenin (DIG)-labelled the DNA probe for northern blotting using the DIG-high prime DNA labelling and detection starter kit I as per the manufacturer's protocol (Roche, catalogue number 11745832910). Northern blots were performed using the DIG northern starter kit as per the manufacturer's protocol (Roche, catalogue number 12039672910), with the following modification: electrophoresis was conducted at 15 V for 24 h at room temperature; RNA was UV-crosslinked (2,400 kJ for 1 min) to a positively charged nylon membrane; gels were blotted by capillary transfer with 20× SSC buffer (3 M NaCl, 0.3 M sodium citrate) overnight; a hybridization temperature of 50 °C was used; blots were hybridized overnight; and 200 µl of NBT/BCIP solution in 10 ml of detection buffer (0.1 M Tris-HCl pH 9.5, 0.1 M NaCl), was used for blot development.

DNA–RNA hybrid immunofluorescence

We seeded 60,000–80,000 cells per PLL-coated coverslip and allowed them to adhere for 24–36 h. Cells were fixed using 1% (*v*/*v*) formaldehyde for 15 min at room temperature, washed three times with 1× PBS for 5 min each, permeabilized with 500 µl of 0.3% (*v*/*v*) Triton-X100 for 5 min at room temperature, and washed again three times with PBS.

Coverslips were blocked using 500 µl of 5% bovine serum albumin (BSA) for 1 h at room temperature, transferred to humidified chambers and incubated with 60 µl of primary antibody (1:500 of S9.6 antibody, 1% (*w*/*v*) BSA, 1× PBS) for 1 h at room temperature. After washing with PBS, cells were incubated with 60 µl of secondary antibody (1% BSA, 1:250 of goat anti-mouse 488 or 1:250 of goat anti-mouse 568) for 1 h at room temperature. The cells were washed again with PBS and incubated with 100 µl of DAPI (0.5 µl of DAPI per ml of PBS) for 2–4 min. The coverslips were then mounted onto microscope slides using DAKO mounting reagent, sealed with nail polish, and allowed to dry for 30 min. Images were acquired using a C2+ confocal microscope with a Plan-Apochromat TIRF ×100 oil objective (numerical aperture 1.45) and NIS-Elements AR software (Nikon). The specificity of the S9.6 antibody for RNA–DNA hybrids was confirmed by in vitro treatment with RNase H1 (NEB, catalogue number M0297S) under the same experimental conditions. Signals were also confirmed to differ from those yielded by immunofluorescence using J2, an antibody against double-stranded (ds)RNA.

Amyloid-body staining with Congo red

We seeded 40,000 cells on PLL-coated coverslips and allowed them to adhere for 24–36 h. Cells were fixed using 1% (*v*/*v*) formaldehyde and incubated for 15 min at room temperature, washed three times with 1× PBS for 5 min each, and permeabilized with 500 µl of 0.3% (*v*/*v*) Triton-X100 for 5 min at room temperature. The coverslips were then immersed in 250 µl of 0.05% (*v*/*v*) Congo red (Sigma, catalogue number C6277) solution for 15 min, followed by four cycles of 2 min rinsing with 500 µl of double-distilled H_2O. The coverslips were then transferred to humidified chambers and nuclear counterstained with 100 µl DAPI, incubated for 4 min, and mounted on microscope slides using DAKO mounting reagent. Images were acquired using a C2+ confocal microscope with a Plan-Apochromat TIRF ×100 oil objective (numerical aperture 1.45) and NIS-Elements AR software (Nikon).

Endogenous protein immunofluorescence

We seeded 40,000 cells onto PLL-coated coverslips. Cells were fixed using 1% formaldehyde for 1 min at room temperature, washed with 1× PBS three times (5 min each wash), permeabilized with 500 µl of 0.3% Triton-X100 for 5 min at room temperature, and washed again three times with 1× PBS. Coverslips were blocked using 500 µl of 5% BSA for 1 h at room temperature, transferred to humidified chambers and incubated with 60 µl of primary antibody (1% BSA and anti-UBF or anti-NPM antibodies) for 1 h at room temperature. After washing with PBS, cells were incubated with 60 µl of secondary antibody (1% BSA, 1:250 of goat anti-mouse 488, 1:250 of goat anti-rabbit 568) for 1 h at room temperature. Coverslips were washed again with PBS, incubated with 100 µl DAPI for 2 min, mounted onto microscope slides using DAKO mounting reagent, and allowed to dry for 30 min. Images were captured at ×100 or ×60 using a Nikon C2+ confocal microscope coupled to NIS-Elements AR software (Nikon). For methanol/acetone-fixation-based immunofluorescence, the protocol was similar except that cells were fixed using ice-cold methanol for 15 min at room temperature, washed once with cold acetone, and washed with 1× PBS (3 × 5 min) before blocking with 5% BSA. Super-resolution microscopy was captured with a Leica DMI6000 SP8 LIGHTNING microscope using the HC PL APO CS2 x93 objective (numerical aperture 1.3, pinhole 110.5 µm). Images were deconvolved using Leica LIGHTNING deconvolution software and processed with Leica LAS software.

Stress-induced droplets and amyloid bodies

The ACM-containing VHL protein was transfected as pFLAG–VHL–GFP using Lipofectamine3000 according to the manufacturer's protocol and treated/visualized 24 h post-transfection. For siRNAs (100 pmol), cells were transfected using RNAiMAX (ThermoFisher Scientific) at 90% confluency, split next day into fresh plates at 70% confluency to allow for subsequent GFP transfection with Lipofectamine3000, and treated/harvested 48 h post-transfection. The siRNAs (ThermoFischer Scientific) used were siControl (catalogue number 4390843), si-sincRNA16 (catalogue number 4399666) and si-sincRNA22 (catalogue number 4390828). For live-cell imaging of the ACM-containing and GFP-tagged VHL protein, cells were seeded in 145-µm-thick, 35-mm glass-bottom plates. Live-cell images were captured by confocal microscopy (Leica TCS SP5; Leica Microsystems, Mannheim, Germany), fitted with a variable temperature and 5% CO_2 environmental chamber (Okolab), using a ×63 oil-immersion Plan-Apochromat objective (numerical aperture 1.4). Images were uniformly adjusted to increase brightness/contrast in Photoshop (Adobe).

In vitro droplet formation

Peptides were custom synthesized by GenScript (New Jersey, USA) at more than 95% purity. Peptide stock solutions were kept at 50 mM in nuclease-free-water. 5′-carboxyfluorescein (FAM)-labelled RNAs were synthesized by Integrated DNA Technologies (IDT, Coralville, IA) and resuspended in 50 mM NaCl to 100 µM. We mixed 1 µM low- or high-complexity ncRNA with the indicated peptide concentrations in 150 mM NaCl. Droplets were placed on a 1.5 coverslip and imaged after a 10-min incubation on a Zeiss AxioObserver D1 microscope using a ×63 Plan-Apochromat objective (numerical aperture 1.4).

Locked nucleic acid ASO knockdown of sincRNAs

Custom-designed locked nucleic acid (LNA) ASO GapmeRs were ordered from Qiagen. Sequences of 975–1,000 bp corresponding to IGS regions were entered into Qiagen's custom antisense LNA GapmeR design page. The top-ranked ASOs based on Qiagen's optimal design score were selected for each of IGS18, IGS20, IGS22 and IGS24, with standard desalting purification, phosphorothioate backbone modifications, and no-label/ready-to-label design specifications. ASO transfection was performed using RNAiMAX (ThermoFisher Scientific) according to the manufacturer's protocols. ASOs were as follows: antisense LNA GapmeR control negative control B (catalogue number 339515, LG0000001-DDA; gctcccttcaatccaa), IGS18 (LG00210930-DDA; agtgtgctctgtgaac), IGS20 (LG00210936-DDA; acgcaagaaaggaaga), IGS22 (LG00210956-DDA; acgtgaccgagagaaa) and IGS24 (LG00210966-DDA; gtgacgtgtagagatt).

Subcellular fractionation by sucrose gradient

Cells were trypsinized, centrifuged at 1,000 r.p.m. for 4 min at 4 °C, washed with PBS and recentrifuged. The pellet was resuspended in osmotic buffer (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT). The cells were then lysed using ten strokes of a tight pestle in a dounce homogenizer. Dounced cells were centrifuged at 1,000 r.p.m. for 5 min at 4 °C. The supernatant was retained as the cytosolic fraction. The pellet was resuspended in a 0.25 M sucrose plus 10 mM $MgCl_2$ solution, and deposited over a 0.35 M sucrose plus 0.5 mM $MgCl_2$ layer. The sample was centrifuged at 1,000 r.p.m. for 5 min at 4 °C. The sample was then resuspended in a 0.25 M sucrose plus 10 mM $MgCl_2$ solution and sonicated at 25% power six times for 10 s with intermittent periods of 10 s rest on ice. The sample was deposited over a 0.88 M sucrose plus 0.5 mM $MgCl_2$ layer and centrifuged at 3,500 r.p.m. for 10 min at 4 °C. The supernatant was retained as the nucleoplasmic fraction. The pellet was resuspended in a 0.35 M sucrose plus 0.5 mM $MgCl_2$ solution and centrifuged at 3,500 r.p.m. for 5 min at 4 °C. The pellet was the nucleolar fraction. GAPDH transcripts, which are most abundant in the cytosolic fraction and are depleted from the nucleolar fraction, served as the control.

Aligning sequencing reads to human rDNA IGS

First, we used the Bowtie package to build a version of the human genome assembly hg19 with rDNA sequence; the newly built assembly is 'hg19_plus_rDNA'. The human rDNA sequence FASTA file was obtained as is from the National Center for Biotechnology Information

(NCBI; https://www.ncbi.nlm.nih.gov) under GenBank accession code U13369.1, which refers to the 'Human ribosomal DNA complete repeating unit'. This FASTA file, along with those for chromosomes 1–22, X, Y and M from hg19, obtained from the University of California at Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu), was used to build the new assembly. Next, for testing, we aligned Pol II reads from HeLa cells to this new genome assembly using the Bowtie package aligner. The reads from two replicates were obtained from ENCODE and concatenated. Duplicate reads were removed with the package BBmap and its clumpify tool. Then, the alignment was performed with the parameter '-m 1', which instructs bowtie to refrain from reporting any alignments for reads having more than one reportable alignment. This ensures that only uniquely aligning reads are reported. The alignments were processed further with Samtools to retain only those reads aligning to the rDNA sequence, and to compute the depth/number of reads at each position in the rDNA coordinates. These depths were plotted with an R script.

Calculation of GC skew

Using the roughly 43-kbp rDNA sequence obtained from the rDNA sequence FASTA file, we assessed GC skew, GC observed/expected ratio and GC percentage using (1-bp-at-a-time) sliding windows of size 50, 500 or 1,000 bp. Definitions are as follows: GC skew = (number of Gs − number of Cs)/(number of Gs + number of Cs); CG observed/expected ratio = sliding-window length × number of CpGs/(number of Cs × number of Gs); GC percentage = 100 × (number of Gs + number of Cs)/sliding-window length. To obtain an overall value/quantification and statistic with which to compare coding and IGS regions, we obtained the mean GC skews for the coding and IGS regions with window size 1,000. In the coding region the mean GC skew is 0.02346459; and in IGS regions, it is −0.1541796. Applying a Welch's *t*-test to the GC skews from these two regions gives a *P*-value of less than 2.2×10^{-16}. The script for all above analyses is called getGCskewEtc_rRDNA.R and is available upon request.

Sequencing

For Ewing- and osteosarcoma-related analyses, sample preparation and sequencing were carried out as described[26]. RNA-seq and DRIP–seq data sets have been deposited at Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession code GSE68847. Identification of rDNA IGS peaks from RNA-seq and DRIP–seq was conducted as per the section 'Aligning sequencing reads to human rDNA IGS' above, including a normalization of called peaks to the total number of reads per sample. For assessment of signals at non-rDNA loci, aligned .bam files were depth normalized and binned using bamCoverage from deepTools[36]. The resulting bigWig files were loaded into Integrated Genome Viewer (IGV)[37] and the depicted regions were saved. For RNA-seq with/without heat shock, sequencing was performed on a cDNA library of total RNA (non-rRNA-depleted) using stranded paired-end reads. After discarding reads mapped to the rRNA gene (including the 5′-ETS, inverted repeat sequence (IRS)1/2 and 3′-ETS), we mapped the remaining reads to GRCh38. BAM files were separated into forward and reverse strand files (bash script). The remaining reads aligned to supercontig GL000220.1; this is within the latest human genome assembly, which contains a 43-kb rDNA cassette. Signals were normalized to an internal non-stress responsive control site at IGS35. To calculate changes in sincRNA and asincRNA levels following heat shock, we binned the IGSs into 5,000-bp bins, and calculated the change in absolute read counts for each bin. The average of these changes was calculated to obtain a global percentage change across the entire IGS region. The sequencing data are available at GEO under accession code GSE115731. ChIP–seq enrichments were generated by the ENCODE Project Broad Institute for H3K27ac, H3K9ac, H3K4me3 and H3K36me3, and by ENCODE Project SYDH for RNA pol II ChIP–seq. Briefly, bedGraph files previously generated[38] by mapping ChIP–seq and input data from ENCODE Project Consortium 2012 to the human rDNA sequence from BAC clone GL000220.1 were used to generate IGV genome tracks. We note that qPCR and sequencing analyses of repetitive DNA loci reveal an average profile for the studied repeats and should not be interpreted as an absolute enrichment for any given unit within the repeats.

Transmission electron microscopy

Cell pellets were fixed in phosphate-buffered 4% formaldehyde plus 1% glutaraldehyde fixative for at least 2 h. Samples were subsequently rinsed in 0.1 M phosphate buffer for 5 min and then fixed in 1% Zetterqvist's buffered osmium tetroxide for 1 h. After a short rinse in Zetterqvist's buffer for 1 min, the samples were dehydrated in increasing concentrations of alcohol (70%, 95%, 100%) for 10 min followed by propylene oxide. Finally, pellets were embedded in epoxy resin. Ultrathin sections were contrasted with uranyl acetate and Reynold's lead citrate and observed with a JEOL 1230 TEM equipped with an Advanced Microscopy Techniques (AMT) camera system.

Images of human tumour sections

Images of tumour sections stained with haematoxylin and eosin were obtained through the Sinai Health System (Toronto) without any identifiable personal health information and without personal information, following Institutional Research Ethics Board approval (Sinai Heath Systems, 17-0103-E).

Statistical analysis

GraphPad Prism-based calculations of *P*-values were carried out via *t*-test, one-way ANOVA (with Dunnett's or Tukey's multiple comparison test), or Mann–Whitney *U*-test. Unless otherwise indicated, replicate information is as follows. All data from pulldowns, reverse transcription and viability markers were generated using the indicated number of biological replicates. For blots, images are representative of data obtained from two independent biological replicates. For microscopy, images are representative of phenotypes observed in at least two independent biological replicates, and quantifications are based on at least 100 cells from two technical replicate cultures.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data are in the paper, Supplementary Fig. 1 (uncropped blots) and the Source Data files related to Figs. 1–4 and Extended Data Figs. 1–3, 5–8. RNA-seq and DRIP–seq data sets have been deposited at GEO under accession codes GSE115731 and GSE68847. All data and materials are available upon reasonable request. In light of the pandemic, shipping of reagents and materials may be slightly delayed. Source data are provided with this paper.

Code availability

All scripts used to analyse data are available upon request.

35. Ho, E. C., Donaldson, M. E. & Saville, B. J. Detection of antisense RNA transcripts by strand-specific RT-PCR. *Methods Mol. Biol.* **630**, 125–138 (2010).
36. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016).
37. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
38. Agrawal, S. & Ganley, A. R. D. The conservation landscape of the human ribosomal RNA gene repeats. *PLoS ONE* **13**, e0207531 (2018).

Acknowledgements We thank D. Durocher and H. O. Lee for critical reading of the manuscript; A. F. Palazzo for technical assistance with phase-contrast microscopy; R. Kandel for technical tumour imaging; and F. Chédin for technical assistance with DRIP. K.J.A. is funded by a

Canadian Institutes of Health Research (CIHR) Vanier Doctoral Scholarship, Ruggles Innovation Award, and Adel S. Sedra Award. N.K. is supported by a CIHR Scholarship. L.A.O. is funded by an Ontario Graduate Scholarship (OGS). R.O. is funded by a Natural Sciences and Engineering Research Council (NSERC) Doctoral Scholarship. This work was also supported by funds from the National Institutes of Health (NIH; grants K22ES012264, 1R01CA152063, 1R01CA241554), a Voelcker Fund Young Investigator Award and the Cancer Prevention and Research Institute of Texas (CPRIT; grant RP150445) to A.J.R.B.; by CPRIT (grant RP101491), a National Cancer Institute (NCI) T32 postdoctoral training grant (T32CA148724) and a National Center for Advancing Translational Sciences (NCATS) TL1 grant (TL1TR002647) to A.G.; and NCI funding (P30CA054174) to the sequencing facility. Funds were also provided to S.L. from the National Institute of General Medical Sciences (R01GM115342) and the NCI (R01CA200676) of the NIH, and the Sylvester Comprehensive Cancer Center. This work was mainly supported by grants to K.M. from the CIHR (388041, 399687), the Canada Research Chairs Program (CRC; 950-230661), and the Ontario Ministry of Research and Innovation (MRI-ERA; ER13-09-111).

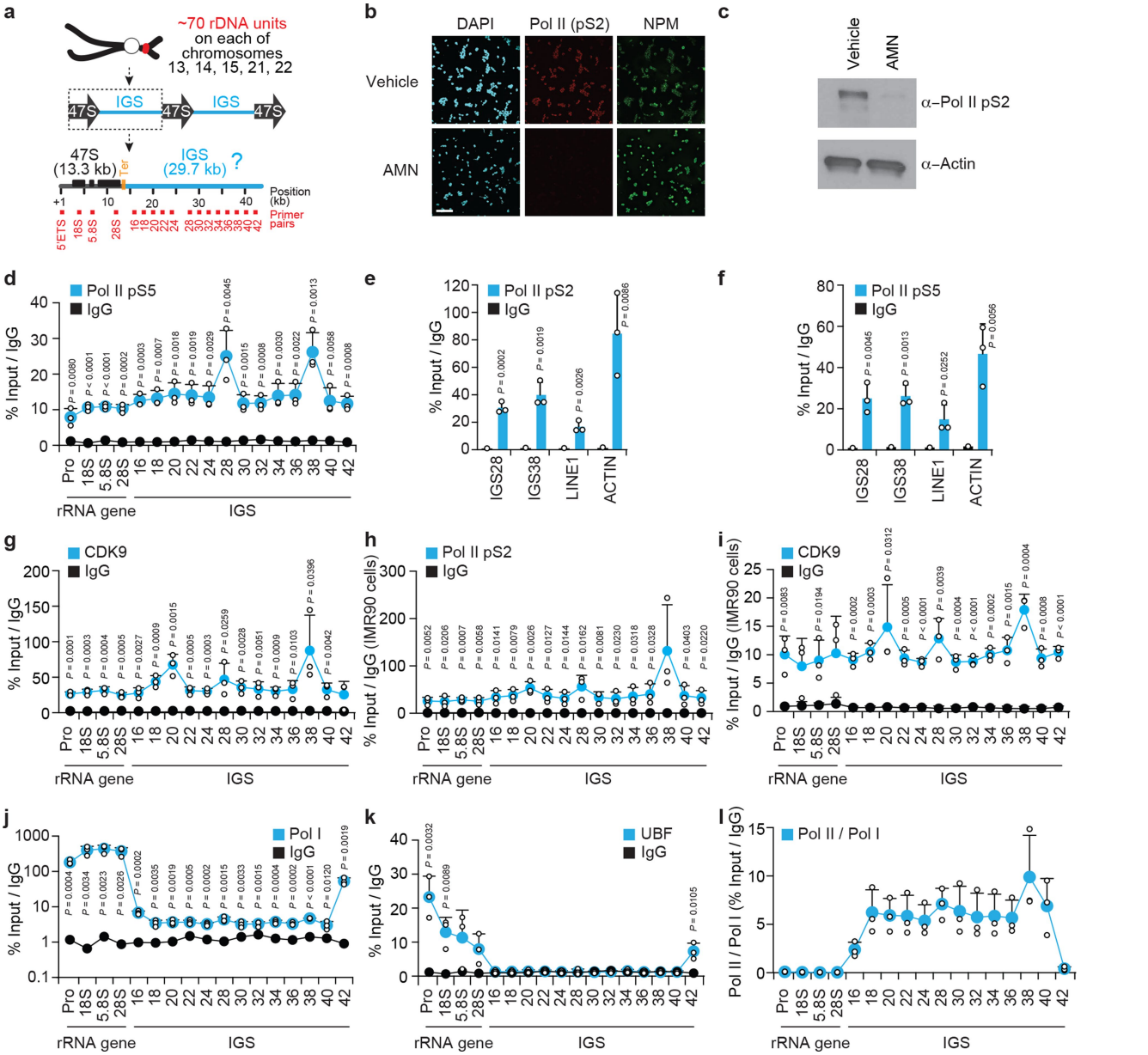
Author contributions K.J.A. and K.M. conceived the study. K.M. together with A.J.R.B. or S.L. conceptualized the Ewing sarcoma and in vitro liquid droplet studies. K.J.A. and K.M. wrote the paper. Most DRIP and ChIP revision experiments and related data analysis were by N.K., supervised by K.J.A. General text editing was by J.N.Y.C., N.K., A.S., L.A.O., R.O., R.H. and J.F.G. The sole lead on all experiments was K.J.A. except as indicated here. Leads were: on 'red laser', K.J.A., N.K. and J.N.Y.C., supervised by K.M.; on ENCODE data analysis, R.S. supervised by D.D.D.C. and V.T.Y. supervised by K.M.; on Ewing sarcoma genome-wide analysis, A.G. supervised by A.J.R.B.; on SETX-knockout generation, D.Y.Z. supervised by J.F.G.; and on heat-shock, cell fractionation and in vitro work, M.W. and M.B. supervised by S.L. DRIP RNH1 overexpression controls were by P.S.P. and A.A., supervised by R.H., and DRIF controls were by V.P., supervised by M.O. and K.M. Tumour sections were by B.C.D. Contributions supervised by K.J.A. and K.M. were from: J.N.Y.C., A.S. and Y.L. for ncRNA detection by northern blotting and RT–qPCR; from E.V. for immunofluorescence microscopy; from L.A.O. for ss-RT design and experimentation; and from R.O. for FRAP design and experimentation.

Competing interests D.D.D.C. received research funds, not related to this work, from Pfizer and Nektar Therapeutics. All other authors declare no competing interests.

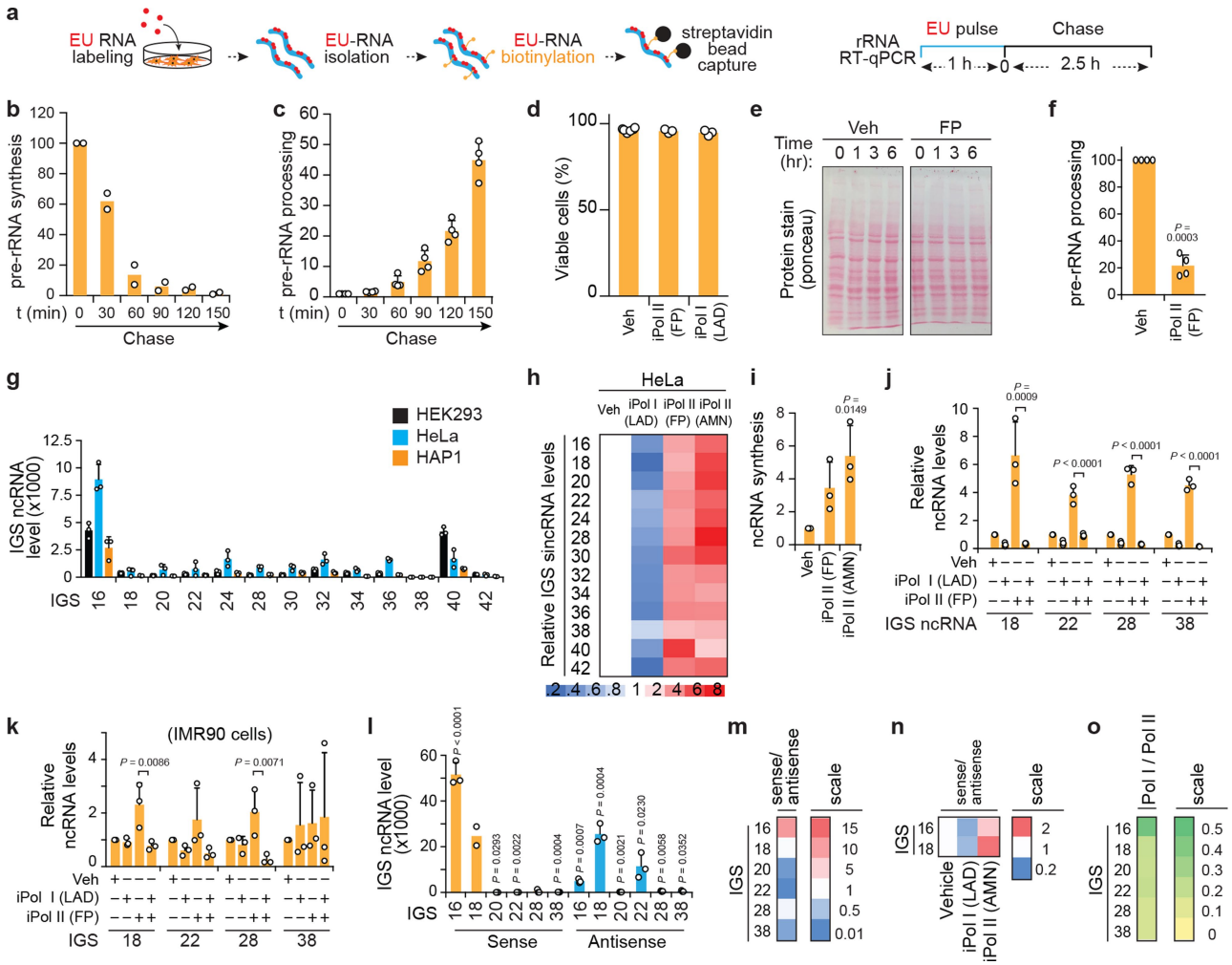
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2497-0.

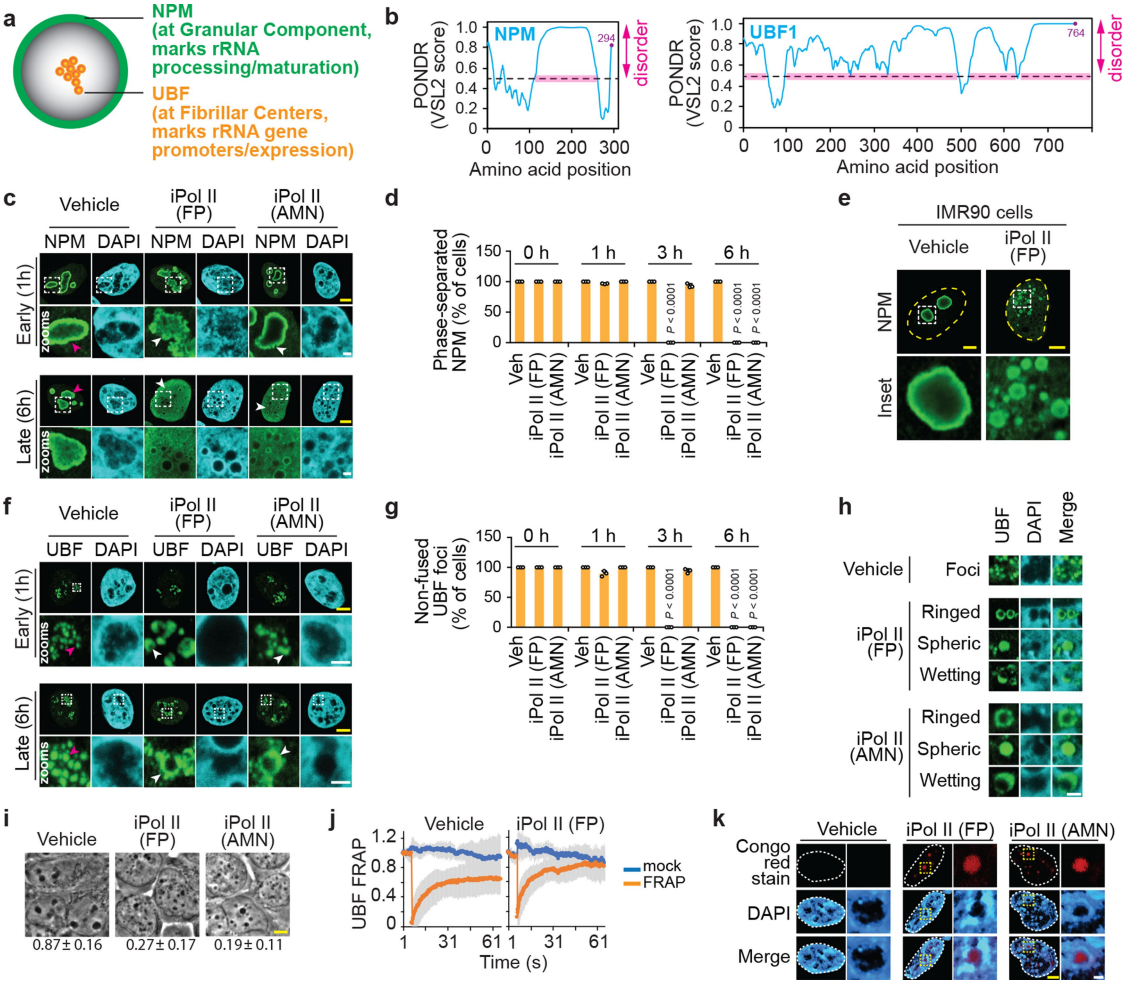
Correspondence and requests for materials should be addressed to K.M.



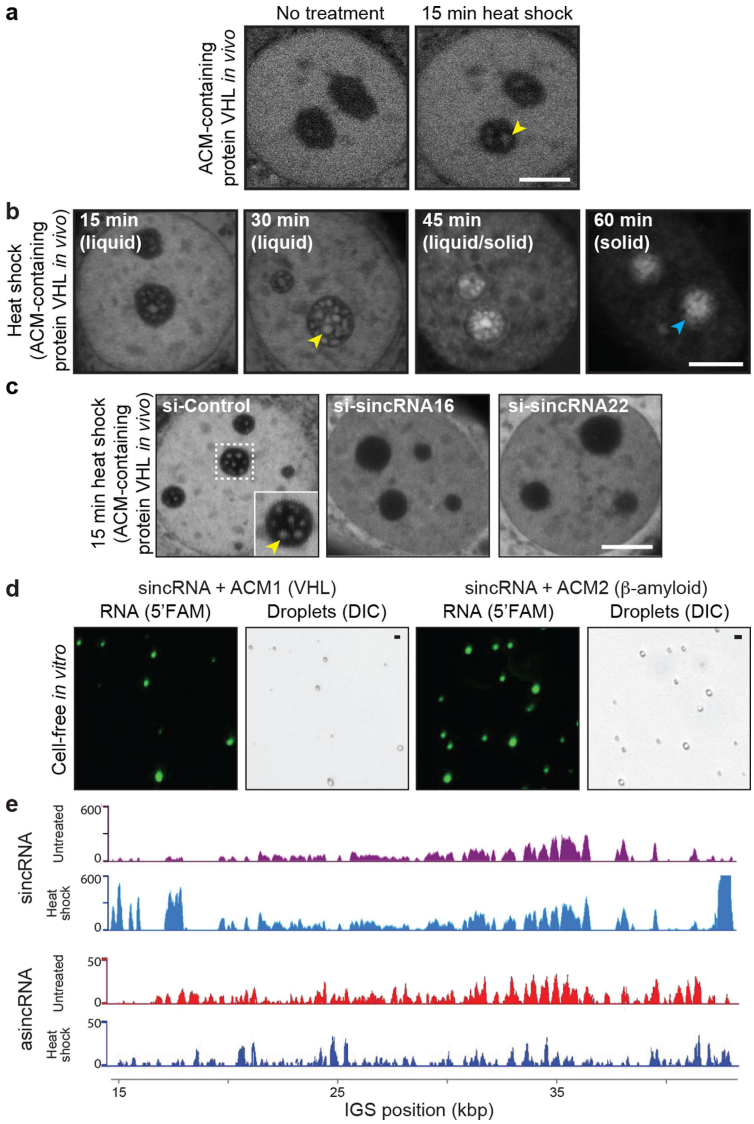
Extended Data Fig. 1 | Additional characterization of Pol I and Pol II occupancy at rDNA IGSs. a, Organization of human rDNA repeats. At each rDNA unit, Pol I transcribes an rRNA gene encoding a 47S pre-rRNA that is processed to remove transcribed spacers, such as the 5′-ETS, and generate 18S, 5.8S and 28S rRNA molecules. The IGS constitutes the bulk of each rDNA unit. Ter, rRNA gene terminator. **b**, **c**, Specificity controls indicating that targeting Pol II for degradation with a 12-hour α-amanitin (AMN) treatment lowers anti (α)-Pol II pS2 signals in both immunofluorescence (**b**) and immunoblotting (**c**). Actin was used as a control for immunoblotting. For gel source data, see Supplementary Fig. 1. **d**, ChIP showing Pol II pS5 enrichment across rDNA. **e**, **f**, The enrichment of active Pol II pS2 and pS5 at rDNA IGS sites is higher than at LINE1 but lower than at β-actin sites. **g**–**k**, ChIP experiments showing the enrichment of the indicated proteins across rDNA. **l**, Comparison of the enrichment of RNA Pol II and Pol I across rDNA reveals the relative overrepresentation of Pol II across IGSs only. **b**–**l**, HEK293T (**b**–**g**, **j**–**l**) or IMR90 (**h**, **i**) cells were used; data shown are means ± s.d.; two-tailed *t*-test, $n = 3$ biologically independent experiments (**d**–**l**); images in **b**, **c** are representative of two independent experiments. Data in **d**–**f**, **j**–**l** and Fig. 1b were from large experimental sets sharing IgG controls. Data in **h**, **i** were from large experimental sets sharing IgG controls.



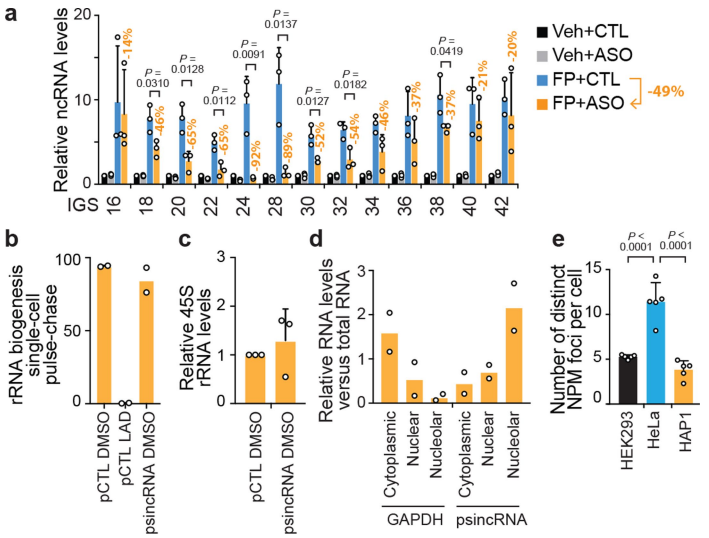
Extended Data Fig. 2 | Impact of Pol I and Pol II on IGS ncRNA levels in various cell types. a, Cell-population-based RNA pulse-chase assay used to assess pre-rRNA synthesis and processing. **b**, **c**, Confirmation of the detection of pre-rRNA synthesis (**b**) and processing (**c**) by EU–RNA pulse-chase assays, as shown in Fig. 1d, e. **d**, Trypan blue exclusion assay confirms that the 3-hour Pol II inhibition (iPol) regimens used in our functional assays do not compromise cell viability. **e**, Ponceau staining shows stable protein levels following Pol II inhibition. Veh, vehicle. **f**, Treatment with the fast-acting RNA Pol II inhibitor flavopiridol (FP) for 30 min is sufficient to abrogate pre-rRNA processing. **g**, Human IGS ncRNAs are also detected across the IGSs of diploid HeLa cells and haploid HAP1 cells. **h**, Pol I promotes and Pol II represses IGS ncRNAs in HeLa cells. **i**, Nuclear run-on assay showing de novo IGS ncRNA synthesis mediated by Pol II inhibition. **j**, **k**, Reverse-transcription experiments showing the effect of combining Pol I and Pol II inhibition on IGS ncRNAs in HEK293T cells (**j**) and IMR90 cells (**k**). **l**, **m**, Strand-specific RT–qPCR (ss-RT) showing the levels of sense and antisense intergenic ncRNAs (**l**) and their derived sense/antisense ratios (**m**) at various IGS sites. **n**, ss-RT shows that Pol I inhibition decreases and Pol II inhibition increases the sense/antisense ratio of the most abundant IGS ncRNAs. **o**, Despite the preferential enrichment of Pol II over Pol I across IGSs, Pol II is the least overrepresented relative to Pol I at IGS16 compared with all other IGSs tested. **a**–**o**, HEK293T cells were used unless otherwise indicated; data are shown as means ± s.d.; two-tailed *t*-test (**b**–**d**, **f**) or one-way ANOVA with Dunnett's multiple comparison test (**g**, **i**, **k**); $n = 2$ biologically independent experiments (**b**), $n = 4$ biologically independent experiments (**c**, **f**), and $n = 3$ biologically independent experiments (**d**, **g**–**o**), except in the case of sense IGS18, for which $n = 2$ biologically independent samples (**l**, **m**); image in **e** is representative of two independent experiments.



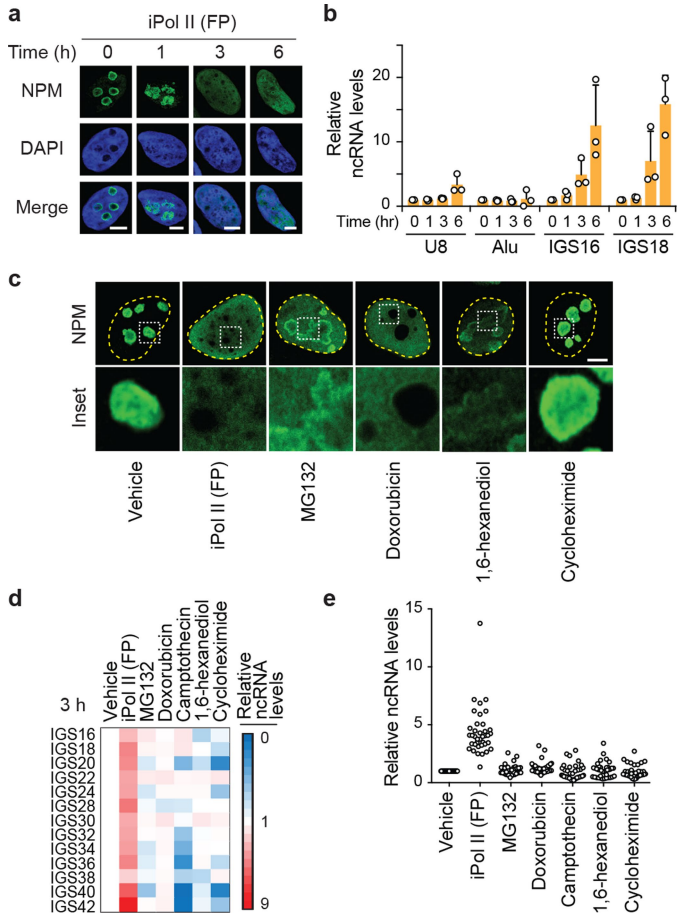
Extended Data Fig. 3 | Characteristics of nucleoli and nucleolar proteins in the presence or absence of Pol II inhibition. a, **b**, Schematic of a nucleolus, illustrating the localization of LLPS nucleolar subcompartments marked by the resident proteins NPM and UBF (**a**), which are highly disordered, as revealed using the various short long 2 (VSL2) predictor of natural disordered regions (PONDR) algorithm (**b**). **c**, Effects of Pol II inhibition (iPol II) on NPM localization, as shown by immunofluorescence microscopy. Examples of normal and defective phenotypes are respectively marked by magenta and white arrowheads. **d**, Quantification of the percentage of cells that have any NPM phase-separated body reveals that the fast-acting Pol II inhibitor FP completely disrupts nucleoli before the slower-acting Pol II inhibitor AMN can take effect. Not depicted is the percentage of cells with perturbed nucleolar architecture as evidenced by NPM1 ruffling, which increased from 0.6 ± 4.6% to 63.3 ± 5.7% following the 1-hour FP treatment. **e**, Pol II inhibition also disrupts NPM localization in IMR90 cells. **f**, Effects of Pol II inhibition on UBF localization, as shown by immunofluorescence microscopy. Examples of normal and defective phenotypes are respectively marked by magenta and white arrowheads. **g**, Quantification of the percentage of cells that have any punctate UBF localization confirmed that the fast-acting FP completely disrupts nucleoli before the slower-acting AMN. **h**, Pol II inhibition triggers various aberrant UBF localization phenotypes, as shown in representative images. **i**, Global nucleolar disruption following Pol II inhibition, as revealed by phase-contrast microscopy. The fraction of cells with more than three black nucleolar bodies is indicated. **j**, Live-cell UBF fluorescence recovery after photobleaching (FRAP). Mock control cells were continuously imaged without a photobleaching step. FRAP FP/vehicle rate-constant ratio = 2.3. **k**, Formerly nucleolar space became Congo red positive after Pol II inhibition. **c**–**k**, HEK293T cells were used unless otherwise indicated; data are means ± s.d.; one-way ANOVA with Dunnett's multiple comparisons test, $n = 3$ biologically independent experiments (**d**, **g**) or $n = 5$ biologically independent experiments (**i**); for **j**, vehicle FRAP cells $n = 30$, vehicle control cells $n = 4$, FP FRAP cells $n = 15$, and FP control cells $n = 6$; images in **e**, **h**, **k** are representative of two independent experiments. Scale bars, 5 µm (yellow) or 1 µm (white).



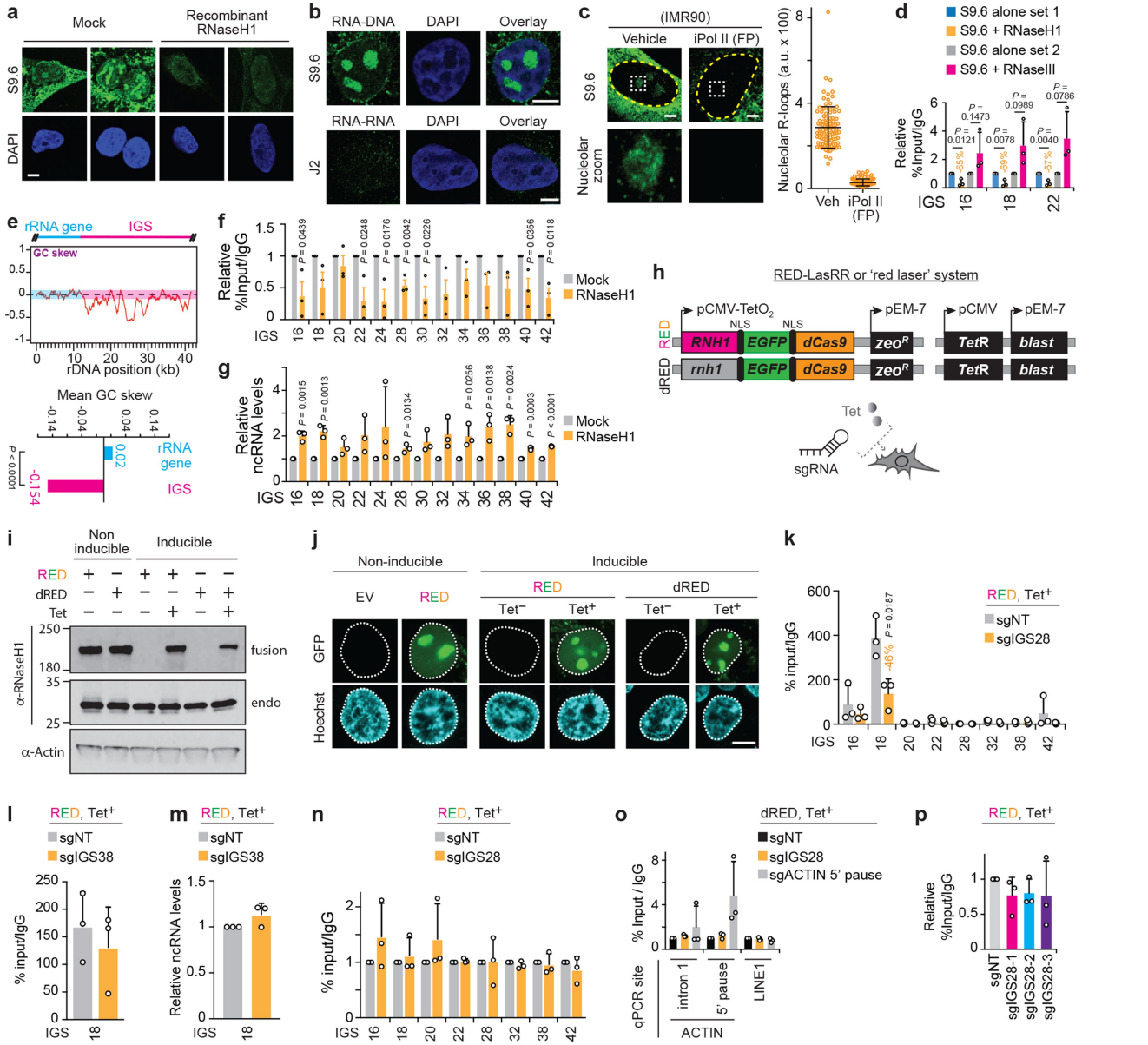
Extended Data Fig. 4 | Heat shock limits asincRNAs and triggers sincRNA-dependent nucleolar phase transitions. a, Heat shock (43 °C) rapidly induces the formation of intranucleolar liquid droplets harbouring the ACM-containing VHL protein. **b**, Gradual amyloid-body (A-body) formation. The stress-induced, mobile and spherical liquid-like foci (yellow arrowhead) gradually transition into irregularly shaped, solid-like amyloid bodies (cyan arrowhead) in cells subjected to heat shock (43 °C)[19]. **c**, The appearance of early-stage, ACM-marked, liquid-like foci[18,19] in cells subjected to a 15-min heat-shock treatment is abrogated upon siRNA-mediated knockdown of either sincRNA16 or sincRNA22. **d**, In a cell-free in vitro system, the low-complexity sincRNA (1 μM) forms liquid droplets when mixed with the ACM of human VHL or β-amyloid proteins (25 μM). Droplets were detected using fluorescently labelled RNA (5′FAM) and differential interference contrast (DIC). **e**, ss-RNA-seq reveals that sincRNA levels increase while asincRNA levels decrease across the IGS following a 30-min heat shock. Heat shock increases sincRNA levels by 607% and decreases asincRNA levels by 38%. **a**–**e**, Nucleolar-stress hyperresponsive MCF7 cells were used where applicable; images are representative of two independent experiments; scale bars, 5 μm.



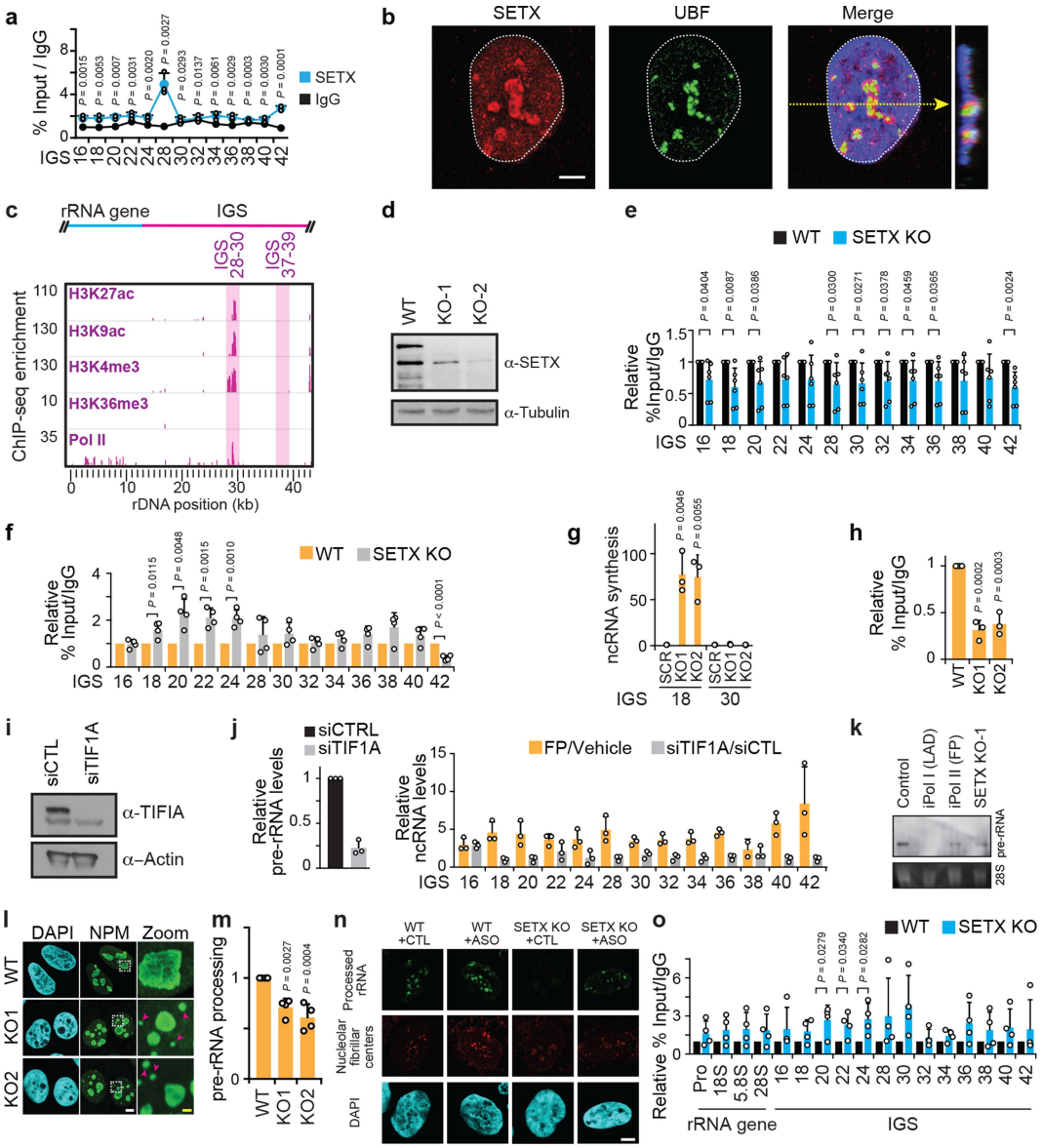
Extended Data Fig. 5 | Artificial and natural modulation of sincRNA levels. **a**, In HEK293T cells treated with the Pol II inhibitor FP, introduction of ASOs targeting sincRNAs lowers IGS ncRNA levels relative to ASO control-treated cells (CTL). ASO-dependent percentage decreases in sincRNA levels are indicated for each IGS site; the average decrease in total sincRNA levels is 49%. Data are means ± s.d.; two-tailed *t*-test, $n = 3$ biologically independent experiments. **b**–**d**, In the absence of heat shock, artificial overexpression of sincRNA22 (psincRNA) in nucleolar-stress hyperresponsive MCF7 cells failed to repress rRNA biogenesis (**b**) or rRNA levels (**c**), despite the enrichment of sincRNA22 in the nucleolar fraction (**d**). Plasmid (pCTL), iPol I (LAD), vehicle (DMSO) and GAPDH cell fractionation controls were included. Data are means ± s.d.; $n = 2$ biologically independent experiments (**b**, **d**); two-tailed *t*-test, $n = 3$ biologically independent experiments (**c**); **e**, Quantification of the number of distinct NPM foci per cell in different cell types. Data are means ± s.d.; one-way ANOVA with Tukey's multiple comparisons test, $n = 5$ biologically independent experiments.



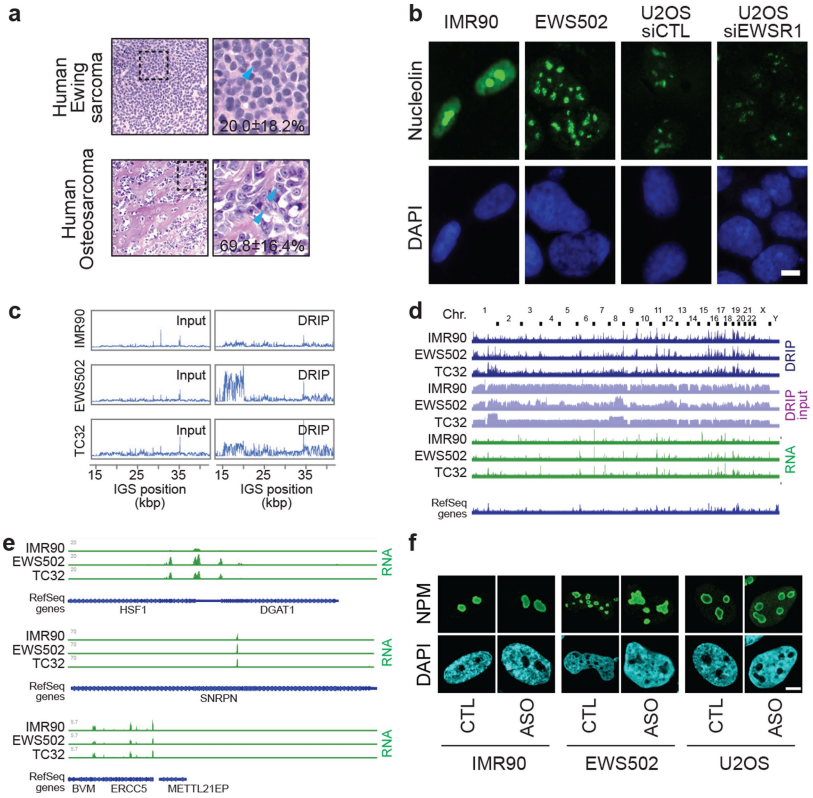
Extended Data Fig. 6 | Controls related to the disruption of nucleolar structure following Pol II inhibition. a, **b**, The disruption of NPM phase separation following Pol II inhibition (**a**) coincides with time points at which the levels of IGS ncRNAs greatly increase (**b**; means ± s.d., $n = 3$ biologically independent experiments). At these time points, no reductions in the levels of the small nucleolar (sno)RNA U8 or Alu RNA were observed. **c**–**e**, Treating cells with the Pol II inhibitor FP, with various drugs that disrupt nucleolar morphology through unclear mechanisms (MG132, doxorubicin), with the LLPS/nucleolus disruptor 1,6-hexanediol, or with the global translation inhibitor cycloheximide reveals that only Pol II inhibition simultaneously disrupted NPM phase separation (**c**) and induced IGS ncRNA levels (**d**, **e**). Shown are representative anti-NPM immunofluorescence images (**c**) and two different visual representations of ncRNA levels as detected by RT–qPCR (**d**, **e**); $n = 3$ biologically independent experiments. In the scatter plot (**e**), each circle represents the value of one IGS site from one of three biological replicates. Scale bars, 5 µm.



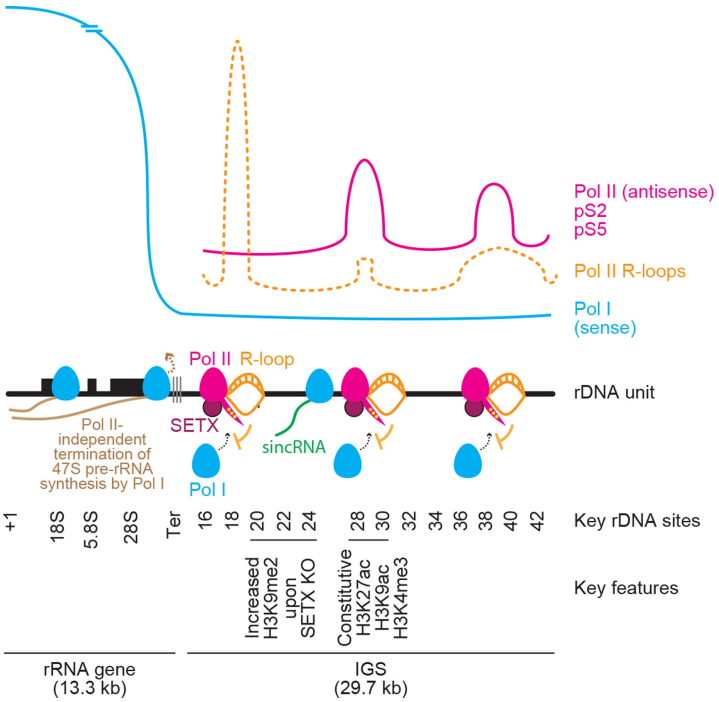
Extended Data Fig. 7 | Nucleolar R-loops and their modulation. a, In vitro treatment with recombinant RNase H1 greatly decreases the nuclear immunofluorescence signals obtained with S9.6, an antibody against DNA–RNA hybrids. Signals remaining following RNase H1 treatment may reflect resistant hybrid structures or other nucleic acid structures. **b**, Immunofluorescence using S9.6, but not the anti-dsRNA antibody J2, reveals a nucleolar signal under standard cell culture conditions. **c**, Immunofluorescence using S9.6 with IMR90 cells also shows nucleolar signals that are repressed upon Pol II inhibition ($n = 100$ cells). **d**, In our DRIP assays, in vitro treatment with RNase H1, but not RNase III, consistently lowers DRIP signals. **e**, Bioinformatic analysis of the rDNA GC skew distribution and mean shows that the IGSs, but not rRNA genes, display a strongly negative GC skew; Welch's two-tailed *t*-test, $n = 14$ (rRNA gene) and $n = 30$ (IGS). **f**, **g**, RNase H1 overexpression partly lowers R-loop levels (**f**) and increases ncRNA levels (**g**) at the IGS. **h**, Design details for the RED/dRED–LaSRR systems created to achieve inducible locus-associated R-loop repression. The zeocin resistance gene (*zeo*^{R}) was used for stable cell line generation, and the blasticidin-resistance gene (*blast*) for selection of the tetracycline repressor (TetR). **i**, **j**, Validation of noninducible and tetracycline-inducible RED and dRED protein expression using immunoblotting (**i**) and microscopy (**j**). For gel source data, see Supplementary Fig. 1. **k**, Using RED together with sgIGS28 decreases R-loop levels at IGS18. **l**, **m**, Using RED together with sgIGS38 fails to alter R-loop (**l**) or ncRNA levels (**m**) at IGS18. **n**, Using RED together with sgIGS28 does not alter Pol II enrichments across the IGS. **o**, The fusion-protein system can be used to preferentially enrich the dRED fusion protein at the 5′ pause site of the *ACTIN* locus. **p**, Use of the nonoverlapping sgRNAs targeting IGS28, individually instead of as a pool, failed to significantly repress R-loop levels at IGS18, arguing against nonspecific effects related to the RNase H1 moiety of RED or any of the gRNAs used. **a**–**p**, HEK293T cells were used unless otherwise indicated. Data are means ± s.d.; one-way ANOVA with Dunnett's multiple comparisons test (**p**; $n = 3$ biologically independent experiments) or two-tailed *t*-test (**d**, **f**–**g**, **k**, **l**–**n**; $n = 3$ biologically independent experiments); $n = 3$ biologically independent experiments (**o**); images in **a**, **b**, **i**–**j** are representative of two independent experiments. Scale bars, 5 µm.



Extended Data Fig. 8 | Nucleolar and IGS features of wild-type and SETX-knockout cells. a, ChIP showing SETX enrichment at the IGS. **b**, SETX has a nucleolar/nucleoplasmic localization. **c**, Bioinformatic analysis of ENCODE-K562 data, showing coenrichment of epigenetic marks consistent with transcriptional activation near IGS28. **d**, Immunoblot showing CRISPR/Cas9-mediated SETX knockout (KO). **e**, ChIP showing Pol II enrichment across rDNA in wild-type and SETX-KO cells. **f**, ChIP reveals that SETX KO, in two clones, enriches RNA Pol I at the IGSs. **g**, **h**, SETX KO induces IGS ncRNA synthesis (**g**) and decreases Pol I enrichment at the rRNA gene (5′-ETS region) (**h**). **i**, **j**, siRNA-mediated knockdown of TIF1A lowers Pol-I-dependent pre-rRNA levels but fails to induce IGS ncRNAs. Because of differences in experimental design, FP/vehicle data (**j**) were from a different experiment (Extended Data Fig. 6d) but are shown here for better visual comparison. **k**, Northern blotting reveals that Pol II or SETX disruption does not induce rRNA gene read-through transcripts. A probe for the 5′-ETS of pre-rRNA was used. **l**, **m**, SETX KO disrupts nucleolar organization as indicated by NPM immunofluorescence (**l**), and decreases pre-rRNA processing in pulse-chase assays (**m**). **n**, ASO-mediated knockdown of sincRNAs increases rRNA biogenesis, as indicated by single-cell rRNA biogenesis assays. Shown are nucleolar fibrillar-centre-associated RNA rings revealed by single-cell FU-RNA pulse-chase immunofluorescence. Quantification shown in Fig. 4e. **o**, ChIP showing H3K9me2 enrichment across rDNA in wild-type and SETX-KO cells. **a**–**o**, HEK293T cells were used unless otherwise indicated. Data in **e**, **o** were from large experimental sets sharing IgG controls. Data are means ± s.d.; two-tailed *t*-test, $n = 3$ biologically independent experiments (**a**, **j**), $n = 6$ biologically independent experiments (**e**), and $n = 4$ biologically independent experiments (**f**, **o**); one-way ANOVA with Dunnett's multiple comparisons test, $n = 3$ biologically independent experiments (**g**, **h**) and $n = 4$ biologically independent experiments (**m**); images in **b**–**d**, **k** are representative of two independent experiments. Scale bars, 5 µm. For gel source data (**d**, **i**, **k**), see Supplementary Fig. 1.



Extended Data Fig. 9 | Additional nucleolar organization and sequencing analyses related to Ewing sarcoma. a, Representative tissue sections of human Ewing sarcoma and osteosarcoma (haematoxylin and eosin staining; magnification ×400). Materials were obtained following Institutional Research Ethics Board approval (Sinai Heath Systems, 17-0103-E). The percentages of cells with one or two distinct nucleoli per nucleus are shown. Data are means ± s.d.; per cancer type, $n = 5$ cases (100 cells each); two-tailed t-test $P = 0.0019$. **b**, Ewing sarcoma cells (EWS502 cells) and U2OS cells with siEWSR1 display disrupted nucleoli, as indicated by the nucleolin protein, compared to their respective control IMR90 and U2OS siControl (siCTL) cells. Scale bar, 5 µm. **c**, Ewing sarcoma (EWS502 and TC32) cells showed increased R-loop levels across IGSs in DRIP–seq. **d**, Genome-wide view of sequence read alignments for DRIP–seq and RNA-seq. Chr., chromosome. **e**, IMR90, EWS502 and TC32 cells can exhibit similarities and differences at non-rDNA loci in sequencing read alignments from RNA-seq. **f**, ASOs targeting sincRNAs ameliorate nucleolar organization. Shown are representative images related to the quantifications in Fig. 4h. Images are representative of two independent experiments. Scale bar, 5 µm.



Extended Data Fig. 10 | Detailed model illustrating how nucleolar Pol-II-dependent R-loops shield the IGS from sincRNA synthesis by Pol I. Top and centre, Pol II at rDNA intergenic spacers (IGSs) synthesizes antisense intergenic ncRNAs (asincRNAs) that constitutively engage in R-loops containing DNA–RNA hybrids (orange). Centre, nucleolar Pol II function is promoted by the neurodegeneration-linked SETX protein (purple). Within rRNA genes, the formation of R-loops usually inhibits the function of Pol I, which is subject to Pol II-independent termination. However, disruption of nucleolar Pol II or its R-loops enables the recruitment of Pol I to IGSs. There, Pol I synthesizes sense intergenic ncRNAs (sincRNAs; green) that mimic environmental stress, disrupting nucleolar liquid–liquid phase separation and triggering an aberrant nucleolar liquid-to-solid phase transition. This unscheduled activation of nucleolar stress responses compromises the natural organization of nucleoli, leading to defects in pre-rRNA biogenesis, especially at the processing level. Nucleolar sincRNA levels are naturally elevated in Ewing sarcoma cells, explaining the indistinct nucleoli often seen in this cancer. In the context of Pol II inhibition, SETX loss or Ewing sarcoma, sincRNA repression ameliorates nucleolar organization and rRNA biogenesis.

nature research

Corresponding author(s): Karim Mekhail, Ph.D.

Last updated by author(s): Apr 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

1) We used a pipeline to align sequencing reads to human rDNA IGS, as described [PMID: 21355038]. A version of the human genome build hg19 with rDNA sequence is built using the Bowtie package (version 1.2.2). The newly built genome assembly is called "hg19_plus_rDNA". The human rDNA sequence FASTA file is obtained as is from NCBI. U13369.1 is the GenBank Accession ID and refers to the "Human ribosomal DNA complete repeating unit" as can be seen on NCBI. This FASTA file along with those from Chr 1-22, X, Y & M from hg19 obtained from UCSC are used to build the new assembly. Next, as control, HeLa Pol II reads are first aligned to this new genome assembly using the Bowtie package aligner. The reads from two replicates are obtained from ENCODE and concatenated. Duplicate reads are removed with the package BBmap (version 37.80) and its clumpify tool. Then, the alignment is performed with the parameter "-m 1" that instructs bowtie to refrain from reporting any alignments for reads having more than 1 reportable alignment. This ensures that only uniquely aligning reads are reported. The alignments are processed further with Samtools (version 0.1.19-44428cd) to retain only the reads aligning to the rDNA sequence, and to compute the depth/number of reads at each position in the rDNA coordinates. These depths are plotted with an R script. The R Project for Statistical Computing (version 3.6.1) from CRAN was used for analysis of IGS read counts across samples.

2) Next, we conducted GC skew calculations across rDNA using R software (version 3.4). Using the ~43K bp rDNA sequence obtained from the rDNA sequence FASTA file, GC skew, CG observed/expected ratio and GC% were assessed using (1 bp-at-a-time) sliding windows of size 50, 500 or 1000 bp. Definitions were as follows: GC skew = (number of Gs - number of Cs)/(number of Gs + number of Cs); CG observed/expected ratio = sliding window length * number of CpGs / (number of Cs X number of Gs); GC % = 100*(number of Gs + number of Cs)/sliding window length. To obtain an overall value/quantification and statistic to compare coding and IGS region, the mean GC skews for the coding and IGS regions with window size 1000 were obtained. In the coding region, the mean GC skew is 0.02346459 and in IGS it is -0.1541796. Doing a Welch Two Sample t-test on the GC skews from these two regions gives a p-value < 2.2x10-16. Script for all above analyses is called getGCskewEtc_rRDNA.R and is available upon request.

3) For Ewing sarcoma-related analyses, sample preparation and sequencing were done as described [PMID: 29513652]. RNA-sequencing and DRIP-sequencing data sources were as described in the Methods section. Identification of rDNA IGS peaks from RNA-seq and DRIP-seq were conducted as per the above described pipeline including a normalization of called peaks to the total number of reads per sample. Tool part of deepTools (version 3.3.0) was used to normalize the .bam files for comparison across samples and Integrated Genome Viewer (version 2.8.0) was used for visualization of .bam files (aligned sequence files).

4) For strand-specific RNA-seq with/without heat shock, sequencing was performed on a non-ribosomal RNA depleted cDNA library of total RNA using stranded paired-end reads as described [PMID: 30110628]. After discarding reads mapped to the rRNA gene (including 5'ETS, IRS1/2, and 3'ETS), the remaining reads were mapped to GRCh38. BAM files were separated into forward and reverse strand files (bash script). The remaining reads aligned to supercontig GL000220.1 that is within the latest human genome assembly that contains a 43 kb ribosomal DNA cassette. Signals are normalized to an internal non-stress responsive control site at IGS35 as described [PMID: 30110628]. The sequencing data source was as described in the Methods section.

5) BioRad CFX Manager (version 3.1) was used for the collection of qPCR data for ChIP and DRIP experiments

6) NIS-Elements AR (version 4.10) was used to acquire microscopy images in DRIF, A-body staining, and endogenous protein immunofluorescence experiments

7) Leica TCS SP5 confocal laser scanning microscope, which uses software platform Leica Application Suite AF (advanced fluorescence) version 2.0.2 was used to collect images in stress induced nucleolar droplets and amyloid bodies experiments. Photoshop (version 20.0.4 CC2018) was used uniformly to adjust brightness and contrast of the images.

8) Zeiss AxioObserver D1 microscope, which uses software platform Zen Blue 2.3 was used for in vitro droplet formation experiments

Data analysis

GraphPad Prism (version 7.0e) was used to display data and perform statistical analyses. ImageJ (version 1.52a) was used to quantify single cell RNA pulse chase and phase contrast imaging experiments. MetaMorph analysis software (version 7.10.3) was used to measure signal intensities in FRAP experiments. Adobe Photoshop CS6 (version 13.0 x64) and Adobe Illustrator CS6 (version 16.0.4) were used to prepare figures for publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data are in the Article, Supplementary Fig. 1 (uncropped blots), and the Source Data files related to Figs. 1-4 and extended data Figs. 1-3 and 5-8. All data and materials are available upon reasonable request. In light of the pandemic, shipping of reagents and materials may be slighted delayed.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. All experiments were conducted with cell lines with multiple available biological or technical replicates as specified in the manuscript, based on previous experience with specific experimental setups, and conforming to field standards. For single cell microscopy experiments, cell counts used per experiment reflect numbers routinely used in stringent quantitative cell biological experiments.
Data exclusions	Exclusion criteria were pre-determined based on internal controls and quality control indicators. For example, any experiment requiring transfection was assessed for successful transfection in parallel before inclusion in data analysis.
Replication	Several observations were tested for their generalizabiliy by 1) assessing multiple cell lines and 2) where applicable multiple knockout clones or chemical inhibitors to rule out clone-specific, cell line-specific, and reagent-specific artifacts. Reproducibility was confirmed by using suitable internal controls to establish validity and replication of findings in biological and technical replicates as indicated in the manuscript.
Randomization	Randomization was not part of the experimental design.
Blinding	Blinding was used for the quantification of microscopy images.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

1) Mouse IgG monoclonal
Concentration: 1 mg/mL
Supplier: Sigma-Aldrich/Millipore
Cat #: 12-371
Application: ChIP, DRIP
Dilution/Usage: ChIP (5 ug per sample), DRIP (10 ug per sample)
Lot#: 3307779, 3267938

2) Rabbit IgG polyclonal
Concentration: 1 mg/mL
Supplier: Abcam
Cat#: Ab171870
Application: ChIP
Dilution/Usage: 5 ug per sample
Lot#: GR3228514

3) H3K9me monoclonal
Concentration: 1 mg/mL
Supplier: Abcam
Cat#: mAbcam 1220
Application: ChIP
Dilution/usage: 5 ug per sample
Lot# GR3228498

4) RNase H1 polyclonal
Concentration: 24 ug/150 ul
Supplier: Proteintech
Cat#: 15606-1-AP
Application: ChIP
Dilution/usage: 5 ug per sample
Lot#: 00043690

5) UBF (F-9) monoclonal
Concentration: 200 ug/mL
Supplier: Santa Cruz
Cat#: sc-13125
Application: IF
Dilution/usage: 1:10
Lot#: H0715, 12413

6) Pol I/RPA135 (N-17)
Concentration: 200 ug/mL
Supplier: Santa Cruz
Cat#: sc-17913
Application: ChIP
Dilution/usage: 5 ug per sample
Lot#: F1714

7) NPM/B23 monoclonal
Clone: FC82291
Concentration: 0.5-0.6 mg/mL

Supplier: Sigma-Aldrich
Cat#: B0556
Application: IF
Dilution/usage: 1:250
Lot#: IC52771

8) Senataxin polyclonal A
Concentration: 1mg/mL
Supplier: Bethyl Laboratories
Cat#: A301-104A
Application: ChIP, ChIP-Re-ChIP, WB
Dilution/usage: ChIP (5 ug per sample), ChIP-re-ChIP (5 ug per sample), WB (1:1000)

9) Senataxin polyclonal B
Concentration: 0.54 mg/mL
Supplier: Novus Bio
Cat#: NBP1-94712
Application: IF
Dilution/usage: 1:250
Lot#: A-1

10) GFP
Concentration: 5 mg/mL
Supplier: Abcam
Cat#: Ab290
Application: WB, ChIP
Dilution/usage: WB (1:1000), ChIP (5 ug per sample)
Lot#: GR3196305

11) RNA-DNA hybrid
Clone: S9.6
Concentration: 1 mg/mL
Supplier: prepared in house by Mekhail lab
Source: ATCC hybridoma (Cat# HB-8730, lot#62851141)
Application: DRIF, DRIP, IF
Dilution/usage: DRIF (1:500), DRIP (10 ug per sample), IF (1:500)

12) BrdU
Clone: BU-33
Concentration: 1 mg/mL
Supplier: Sigma-Aldrich
Cat# B2531
Application: single cell pulse chase
Dilution/usage: 1:250
Lot# 038M4861V

13) RNA polymerase II CTD repeat YSPTSPS (pS2)
Concentration: 1 mg/mL
Supplier: Abcam
Cat#: Ab5095
Application: ChIP, WB, IF
Dilution/usage: ChIP (5 ug per sample), WB: (1:1000), IF (1:600)
Lot#: GR3278442, GR3225147, GR3172948, GR231750

14) RNA polymerase II CTD repeat YSPTSPS (pS5)
Concentration: 1 mg/mL
Supplier: Abcam
Cat#: Ab5048
Application: ChIP
Dilution/usage: 5 ug per sample
Lot#: GR205997

15) ATXN2 polyclonal
Concentration: 200 ug/mL

Supplier: Sigma-Aldrich
Cat# HPA021146
Application: IF
Dilution/usage: 1:250
Lot#: A113803

16) dsRNA J2
Clone: rJ2
Concentration: 1 mg/mL
Supplier: Sigma-Aldrich/Millipore
Cat# MABE1134-100UL
Application: DRIF
Dilution/usage: 1:600
Lot#: 3170762

17) RRN3/Tif1A polyclonal
Concentration: 1 mg/mL
Supplier: Abcam
Cat# ab112052
Application: WB
Dilution/usage: 1:1000
Lot#: GR251820

18) Beta-Actin
Clone: mAbGEa
Concentration: 1 mg/mL
Supplier: Invitrogen/Thermo Fisher
Cat# MA1-744
Application: WB
Dilution/usage: 1:1000
Lot#: UB272750

19) CDK9 polyclonal
Supplier: Proteintech
Cat# 11705-1-AP
Application: ChIP
Dilution/usage: 5 ug per sample
Lot# 00047991

Validation

Commercially available antibodies were validated for specificity by the manufacturer using knockdown or knockout of cognate transcript/gene. The SETX antibody was additionally validated for specificity using CRISPR/Cas-mediated knockout of SETX. In addition, the specificity of our S9.6 antibody for RNA-DNA hybrids was validated using in vitro treatment with RNase H1, in vivo over-expression of RNase H1, RED-mediated signal repression, and dRED-mediated signal amplification. RNase H1 controls are also included in individual experiments to ensure that signals reflect RNA-DNA hybrids.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HEK293T, HeLa, HAP1, IMR90, MCF7 and U2OS cell lines were purchased from ATCC. HEK293T T-REx™ cells were purchase from ThermoFisher Scientific. EWS502 cells were from Dr. A. J. R. Bishop, who had previously obtained the cells from Dr. S. Lessnick. TC32 cells were also from Dr. A. J. R. Bishop, who had previously obtained them from the Children's Oncology Group.

Authentication

Purchased cell lines were commercially authenticated by ATCC or ThermoFisher Scientific. Cells obtained from Dr. A. J. R. were previously authenticated [PMID: 29513652]. Specifically, Following sequencing, identity was confirmed using known mutations in the cell lines, in addition to performing STR profiling on TC32 and U2OS. For all cell lines, cutlures were not maintained for more than 6 months prior to returning to low passage stocks.

Mycoplasma contamination

The cell lines used tested negative for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	DRIP-seq, which is similar to ChIP-seq, was used. Raw and processed files used for alignment to rDNA IGS are previously reported [PMID: 29513652] and described in the Methods section.
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	Experiments were previously done with biological replicates (EWS502 and TC32 Ewing cell lines) [PMID: 29513652].
Sequencing depth	The analysis conducted in our paper is based on published data that were obtained from samples that were amplified by PCR through 40 cycles. These samples were then processed as 50bp single-end sequencing and sequenced with 30-50 million reads for each sample.
Antibodies	Antibody against: RNA:DNA hybrids Supplier: Kerafast Catalog Number: ENH001 Clone number: S9.6 Validation: The specificity of the antibody was validated by: (1) References listed on the company website (https://www.kerafast.com/product/1552/anti-dna-rna-hybrid-s96-antibody), (2) Using samples treated with RNaseH1 to demonstrate specificity to mark RNA-DNA hybrids, and (3) qPCR on known R-loop sites as well as sites that are known to not have R-loops.
Peak calling parameters	The alignment is performed with the parameter "-m 1" that instructs bowtie to refrain from reporting any alignments for reads having more than 1 reportable alignment.
Data quality	Data quality was controlled with FDR < 0.05.
Software	We used a published pipeline for aligning sequencing reads to human rDNA IGS [PMID: 21355038]. A version of the human genome build hg19 with rDNA sequence is built using the Bowtie package. The resulting genome assembly is called "hg19_plus_rDNA". The human rDNA sequence FASTA file is obtained as is from NCBI. U13369.1 is the GenBank Accession ID and refers to the "Human ribosomal DNA complete repeating unit" as can be seen on NCBI. This FASTA file along with those from Chr 1-22, X, Y & M from hg19 obtained from UCSC are used to build the new assembly. Next, sequencing reads are aligned to this new genome assembly using the Bowtie package aligner. Duplicate reads are removed with the package BBmap and its clumpify tool. Then, the alignment is performed with the parameter "-m 1" that instructs bowtie to refrain from reporting any alignments for reads having more than 1 reportable alignment. This ensures that only uniquely aligning reads are reported. The alignments are processed further with Samtools to retain only the reads aligning to the rDNA sequence, and to compute the depth/number of reads at each position in the rDNA coordinates. Signals are normalized to the total number of reads per sample. The normalized depths are plotted with an R script.

Article

Shared structural mechanisms of general anaesthetics and benzodiazepines

https://doi.org/10.1038/s41586-020-2654-5

Received: 20 December 2019

Accepted: 1 June 2020

Published online: 2 September 2020

Check for updates

Jeong Joo Kim[1], Anant Gharpure[1], Jinfeng Teng[1], Yuxuan Zhuang[2], Rebecca J. Howard[2], Shaotong Zhu[1], Colleen M. Noviello[1], Richard M. Walsh Jr[3], Erik Lindahl[2,4] & Ryan E. Hibbs[1]✉

Most general anaesthetics and classical benzodiazepine drugs act through positive modulation of γ-aminobutyric acid type A ($GABA_A$) receptors to dampen neuronal activity in the brain[1–5]. However, direct structural information on the mechanisms of general anaesthetics at their physiological receptor sites is lacking. Here we present cryo-electron microscopy structures of $GABA_A$ receptors bound to intravenous anaesthetics, benzodiazepines and inhibitory modulators. These structures were solved in a lipidic environment and are complemented by electrophysiology and molecular dynamics simulations. Structures of $GABA_A$ receptors in complex with the anaesthetics phenobarbital, etomidate and propofol reveal both distinct and common transmembrane binding sites, which are shared in part by the benzodiazepine drug diazepam. Structures in which $GABA_A$ receptors are bound by benzodiazepine-site ligands identify an additional membrane binding site for diazepam and suggest an allosteric mechanism for anaesthetic reversal by flumazenil. This study provides a foundation for understanding how pharmacologically diverse and clinically essential drugs act through overlapping and distinct mechanisms to potentiate inhibitory signalling in the brain.

General anaesthetics were long thought to act through a membrane effect owing to a strong correlation between their potency and their tendency to partition into lipid[6–9]. This non-specific model became harder to reconcile after the discovery of exceptions to the rule, including isomers of anaesthetics with opposing activities[10–12]. More recent electrophysiology[11], mutagenesis[13] and labelling studies[2]—together with mouse knock-in studies[14]—have identified the $GABA_A$ receptor as the principal target for most modern intravenous anaesthetics. The first intravenous anaesthetics were barbiturates, which were developed in the 1930s as anticonvulsive drugs. However, they have a narrow therapeutic index and have been largely replaced by etomidate and propofol, which are more selective and are the two most frequently used intravenous anaesthetics today. Like classical benzodiazepines, all general anaesthetics that act through the $GABA_A$ receptor are positive allosteric modulators; however, the transmembrane sites of general anaesthetics are distinct from those at which benzodiazepines are mainly thought to act.

Benzodiazepines are $GABA_A$ receptor ligands that are used in the treatment of epilepsy, anxiety and insomnia[3,4]. Classical benzodiazepines such as diazepam are positive allosteric modulators of $GABA_A$ receptors and exhibit a range of pharmacological effects, from sedation at low doses to the induction of anaesthesia at higher doses. These different effects have been related to the presence of two distinct classes of binding site on the receptor. A high-affinity benzodiazepine site at the α–γ subunit interface in the extracellular domain of the receptor is responsible for the positive modulation that is useful in treating anxiety and seizure disorders. One or more lower-affinity sites are thought to contribute to the ability of high doses of some benzodiazepines, such as diazepam, to directly induce anaesthesia[15,16]. Flumazenil is a competitive antagonist of the α1-γ2 high-affinity benzodiazepine site[3,5] that is used clinically as an antidote for benzodiazepine overdose and to reverse general anaesthesia[17]. The structural mechanisms that underlie potentiation by benzodiazepines and their antagonism by flumazenil have begun to emerge, but remain largely unclear.

Here we investigate the structural basis of how intravenous anaesthetics modulate $GABA_A$ receptor signalling, and how their mechanisms overlap in part with those of benzodiazepines. We optimized a lipid-reconstitution approach for the α1β2γ2 $GABA_A$ receptor to determine structures in complex with GABA plus the barbiturate phenobarbital, with GABA plus etomidate, and with GABA plus propofol, mapping their distinct binding sites and atomic interactions. We compare these anaesthetic complexes to new structures of the α1β2γ2 $GABA_A$ receptor bound by GABA alone, GABA plus diazepam, and GABA plus flumazenil, to define common and distinct mechanisms for potentiation, and elucidate how flumazenil antagonizes the positive modulators in a competitive or allosteric manner. We then analyse structures of the α1β2γ2 receptor bound by GABA plus picrotoxin (pore blocker) and by bicuculline (competitive antagonist) to enable comparison with recent structures of the highly similar α1β3γ2 receptor[18,19]; we identify systematic conformational differences between these structures, which probably arise from the lipid-reconstitution method. Mutagenesis, electrophysiology and molecular dynamics simulations complement the structural findings on ligand recognition and conformational stabilization.

[1]Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden. [3]Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [4]Department of Applied Physics, Swedish e-Science Research Center, KTH Royal Institute of Technology, Solna, Sweden. ✉e-mail: ryan.hibbs@utsouthwestern.edu

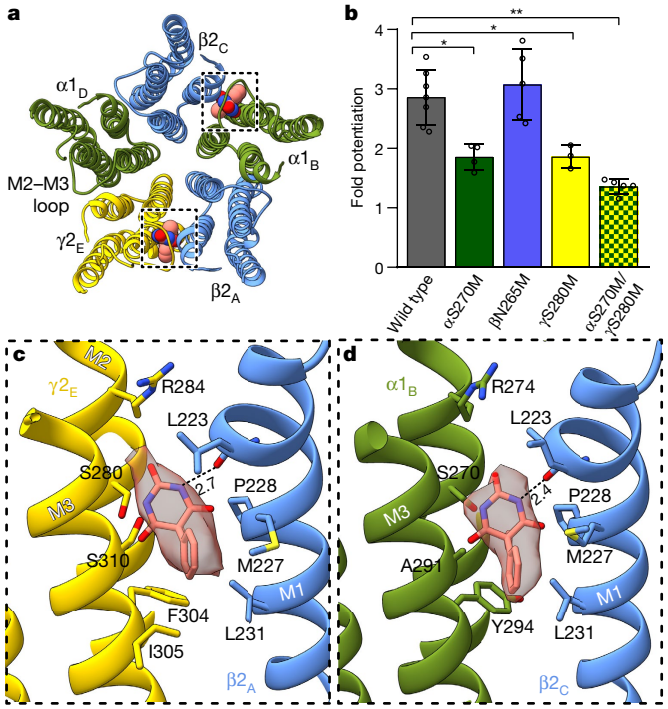


Fig. 1 | Phenobarbital-binding sites. a, The atomic model of TMD viewed down the channel axis from the synaptic perspective. The boxes highlight phenobarbital sites with the ligands shown as spheres. **b**, The effect of mutation at the 15′ position of different subunits on the potentiation of GABA activation by phenobarbital. Data are mean ± s.d. $n = 7$ (wild type), 4 (αS270M), 5 (βN265M), 3 (γS280M) and 5 (αS270M/γS280M) biologically independent patch-clamp experiments with individual cells. $^{*}P < 0.01$; $^{**}P < 0.0001$. **c**, **d**, Details of the binding site of phenobarbital at the γ–β (**c**) and α–β (**d**) interfaces. Hydrogen bonds are indicated with dashed lines, and their distance (in Å) is given.

Barbiturate recognition

Barbiturates exhibit a range of $GABA_A$ receptor-mediated activities, including sedative, anxiolytic, hypnotic and anticonvulsant effects. Phenobarbital in particular remains popular as an antiepileptic drug. At low concentrations it potentiates the response of the receptor to GABA, whereas at high concentrations it evokes direct allosteric activation, through binding sites in the transmembrane domain[20]. We developed a lipid-reconstitution approach (Extended Data Figs. 1, 2, Methods) to stabilize the transmembrane domain (TMD) and avoid the collapse of the pore that occurs when using detergent[21], then collected cryo-electron microscopy (cryo-EM) data on the α1β2γ2 receptor in complex with GABA plus phenobarbital. Despite approximate five-fold symmetry in the membrane domain for the barbiturate complex, density for the TMD of the γ2 subunit (γ-TMD) was weaker than for other subunits—an observation common to all ligand complexes we studied—with functional implications that are potentially relevant to desensitization[22]. We therefore performed focused 3D classification on the γ-TMD to improve the local signal; this resulted in a 3.1 Å resolution map with strong signal in the γ-TMD and with clear density for phenobarbital at the α–β and the γ–β interfaces (Fig. 1, Extended Data Figs. 3, 4, Extended Data Table 1).

The binding mode of phenobarbital is equivalent for the two sites: the barbituric acid group rests deep in the M3–M1 interfaces, the phenyl group orients away from the channel axis and the ethyl group orients towards the pore (Fig. 1c, d, Supplementary Videos 1, 2). The binding locus is at the level of the M2 15′ residue and is just below a short π-helix in β2 M1. Notably, the proline residue responsible for this π-helix is conserved across diverse members of the Cys-loop receptor superfamily; it creates a bulge in M1 that in turn creates a pocket that is present at all interfaces in the structure below the M2–M3 loop. Phenobarbital is stabilized mainly through van der Waals interactions, and forms one electrostatic contact between the backbone carbonyl oxygen of βL223 and a barbituric acid nitrogen. Although these two binding modes are consistent in position and in predicted pose with the results of a previous affinity-labelling analysis[23], they are in potential conflict with a study in β3 point-mutant knock-in mice that predicts binding at β–α interfaces[24]. Mice with an asparagine-to-methionine mutation at residue 265 (N265M) of β3—the M2 15′ residue that would contribute to a β–α interface site—exhibit a partial loss of anaesthetic response to pentobarbital. This β2/3 15′ asparagine residue corresponds to S270 in α1 and S280 in γ2 (Fig. 1c, d). We observed no density for phenobarbital at the β–α interface and, based on the structure, substitution of the 15′ serine with asparagine would result in a steric clash with phenobarbital. To assess the roles of the γ–β, α–β and potential β–α interfaces in the sensitivity to barbiturates in vitro, we mutated the residues at these homologous positions to methionine—which renders receptors less sensitive or unresponsive to anaesthetics[25]—and tested the sensitivity of these mutant receptors to potentiation of low-dose GABA with phenobarbital (Fig. 1b). Individually mutating the residues at each interface at which we modelled phenobarbital resulted in a marked loss of potentiation, which increased in the double mutant. By contrast, mutation at the β265 position resulted in no notable difference in potentiation by phenobarbital. The results of mutagenesis, electrophysiology, structural biology and affinity-labelling studies are thus internally consistent in defining two important barbiturate sites in the $GABA_A$ receptor TMD: at the α–β and γ–β interfaces.

Recognition of etomidate and propofol

Propofol is the most widely used intravenous general anaesthetic[26]. Etomidate preceded propofol in development and is currently used instead of propofol in cases in which cardiovascular or respiratory depression is a concern[26,27]. In vitro mutagenesis studies identified binding sites for both compounds at the β–α interfaces, in positions equivalent to the β–α TMD binding sites of diazepam; these binding sites were responsible for the potentiation of GABA binding by these compounds, as well as for the direct activation of the receptor at higher concentrations. Evidence for etomidate binding exclusively at the β–α interfaces is strong[28–30]. Affinity-labelling and mutagenesis studies of propofol suggest that, in addition to the β–α sites[31–33], there may be additional binding sites at other subunit interfaces[32,34] and/or at the TMD–ECD (extracellular domain) junction[35]. Mouse knock-in studies of mutated $GABA_A$ subunits connected the immobilizing and the sedative/hypnotic effects of both drugs to the β–α TMD sites in receptors containing β3[36] and β2[37] subunits, respectively, with a caveat that the results were less clear in the β2 knock-in mice for propofol than for etomidate. We obtained structures of the α1β2γ2 $GABA_A$ receptor in complex with etomidate (3.5 Å resolution) and propofol (2.6 Å resolution) in order to directly interrogate binding interactions and provide a foundation for understanding allosteric potentiation.

Both the etomidate and propofol density maps revealed clear signal for ligands at β–α interfaces at the predicted binding sites, and no corresponding density at the other interfaces (Fig. 2a–d, Extended Data Figs. 3, 4, Extended Data Table 1, Supplementary Videos 3, 4). The pose for both ligands, at each of the two β–α interfaces, is equivalent. Etomidate binds at the β–α interfaces at the same level as phenobarbital, with its phenyl ring orienting towards the cytosol, its methyl and imidazole groups orienting towards the channel axis and its ethyl ester orienting away from the pore, towards bulk lipid (Fig. 2c, Supplementary Video 3). The orientation of etomidate is markedly similar to that predicted from affinity-labelling studies[38]. Its phenyl ring packs against N265 of β15′, probably forming an electrostatic interaction between the amide nitrogen of the side chain and the π electrons of the

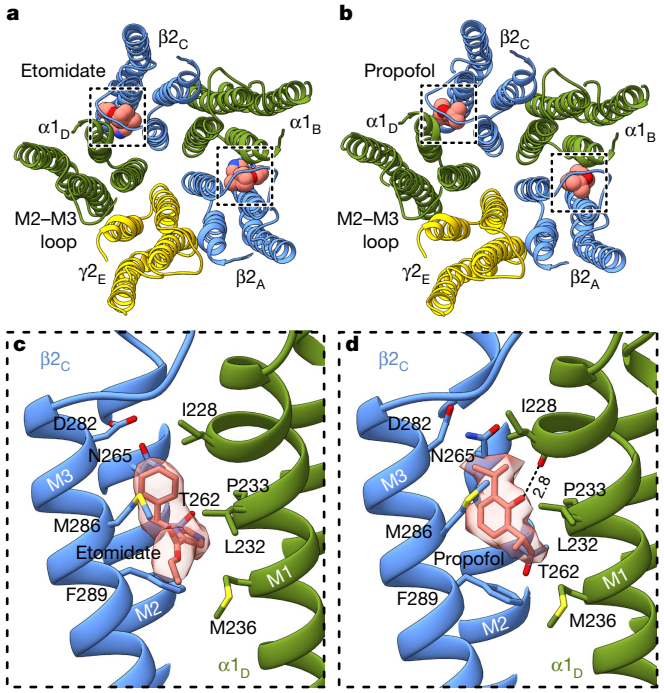


Fig. 2 | Interactions of etomidate and propofol. a, **b**, Atomic model overview of the TMD binding sites for etomidate (**a**) and propofol (**b**); ligands are shown as spheres. The subscripts of the subunits identify the chains. **c**, **d**, Details of the binding site of one of the two equivalent β–α sites for each ligand. Experimental density for ligands is shown as a semi-transparent surface.

phenyl ring. In a related receptor assembly, mutation of N265 at this β15′ position to serine resulted in a tenfold loss of etomidate sensitivity, whereas mutation to methionine resulted in total loss of etomidate potentiation[28]. The imidazole ring of etomidate is sandwiched between βF289 in the M3 helix and αP233 across the interface in M2. Extensive van der Waals contacts are made at the interface, including with the side chain of βM286. Mutation of M286 to tryptophan results in a large loss of sensitivity to etomidate[29], which is consistent with all common rotamers of tryptophan in this position generating clashes with either etomidate or the receptor.

The high resolution of the complex of the receptor with GABA plus propofol enabled confident positioning of propofol at both β–α interfaces, in a position overlapping with that of etomidate (Fig. 2d, Supplementary Video 4). Propofol is symmetric and is smaller than etomidate, and makes fewer contacts with the receptor. One isopropyl group orients towards the channel axis and one towards bulk lipid; this latter hydrophobic group packs against the αP233 that creates the M1 π-helix. The channel-proximal isopropyl group orients towards the β15′ position, forming van der Waals contacts. Substitution of this residue with serine has little effect on the response of knock-in mice to propofol[37], which can be rationalized by considering the structure of the complex; unlike etomidate, propofol is not oriented to form electrostatic interactions with the asparagine at the 15′ position. By contrast, knock-in mice harbouring a methionine in the 15′ position of β3 are insensitive to the immobilizing effects of propofol[36]; the long hydrophobic side chain of methionine would compete directly for propofol binding in the structure. The benzyl ring is oriented with its face parallel to the membrane normal; its hydroxyl extension, a hydrogen-bond-donating group known to be a determinant of propofol potency[39], forms a hydrogen bond with the backbone carbonyl oxygen of αI228 that is liberated by the M1 π-helix. βM286 reaches across the subunit interface such that it could limit or slow exchange between the bound propofol and bulk lipid; mutation of this residue to tryptophan causes a loss of propofol potentiation[40], as is also observed for etomidate[29]. Simulations to assess propofol binding at the other three TMD interfaces suggest that it is less stable in those locations (Extended Data Fig. 1h). Although we cannot rule out the possibility of propofol binding at additional sites[41], the structural analysis—combined with mutagenesis, affinity labelling and animal studies—is consistent with high-affinity binding of both etomidate and propofol only at β–α interfaces in the TMD. Notably, at the α–β and the α–γ interfaces, density consistent with a lipid head group occupies the propofol-equivalent position. At the two β–α interfaces, lipid density is also present but is peripheral to the site (Extended Data Fig. 5a–e).

Mechanisms of benzodiazepines

We next relate these insights into anaesthetic recognition to a distinct class of allosteric modulators, the benzodiazepines. To survey a range of activities, we obtained structures of the receptor in complex with GABA plus the benzodiazepine-site-antagonist flumazenil (3.5 Å), with GABA and an apo benzodiazepine site (3.2 Å), and with GABA plus the positive modulator diazepam (2.9 Å; Extended Data Figs. 2–7, Extended Data Table 2). We discuss these three structures in detail in the Supplementary Information, and focus here on new findings and emergent trends. In the complex with flumazenil, we found near-perfect agreement between the ECD of this structure and that of the same complex in detergent[21] (Extended Data Fig. 6f, g). The relatively high disorder in the γ-TMD observed in all structures was most notable in the flumazenil complex, in which a gap is present at the γ–β interface (Fig. 3a–f, Extended Data Fig. 2e, f). This gap shrinks in the absence of flumazenil and disappears in the presence of diazepam. In the diazepam complex, in addition to the expected density for diazepam at the classical benzodiazepine site at the ECD α–γ interface (Extended Data Fig. 7a, b), we observed three distinct densities for diazepam in the transmembrane domain: two at β–α interfaces as observed previously[19], and a third at the γ–β interface that overlaps with one of the phenobarbital sites (Fig. 3c, Extended Data Fig. 7d–f, Supplementary Videos 7, 8). Binding of diazepam to this latter site may contribute to the overall stability of the TMD by closing the γ–β gap, similar to what was observed with the barbiturate, and may also have a role in benzodiazepine-induced potentiation through a mechanism similar to that of anaesthetics[15,16]. In this new class of diazepam-binding site at the γ–β TMD interface (Extended Data Fig. 7d, f, Supplementary Video 8) the diazepine ring pucker inverts, adopting an enantiomeric conformation (Extended Data Fig. 7g). In contrast to the β–α sites, the diazepam at the γ–β interface positions above γS280, homologous to βN265. In this pose, the pendant phenyl ring of diazepam points away from the channel axis and interacts with conserved phenylalanine (γF304) and proline (βP228) residues (Extended Data Fig. 7f, Supplementary Video 8). Consequently, the benzyl ring is located near the γM2 helix and the diazepine carbonyl oxygen forms a hydrogen bond with γT277. Investigation of the other intersubunit sites in the TMD revealed tubular density at the α–β interface, which shares sequence similarity with the γ–β site and has been proposed to be an active binding site for benzodiazepines[15,16] and barbiturates[23], as well as at the α–γ interface. Molecular dynamics simulations suggest that these densities probably correspond to lipids (Extended Data Fig. 5f, g, Supplementary Videos 9, 10). Our structural and simulation results thus support the existence of an orphan site that does not respond to benzodiazepines or to anaesthetics[2].

Occupancy of four sites by diazepam results in global stabilization compared to the complex with GABA alone, and especially compared to the complex with GABA and flumazenil (Fig. 3a–f). Together, these three structures—together with those of the anaesthetic-bound complexes—reveal a correlation between receptor stability in the TMD and activity of the allosteric ligand. In contrast to the stabilization in the TMD that is observed with positive modulator complexes, the binding of flumazenil destabilizes the TMD and results in a slightly expanded

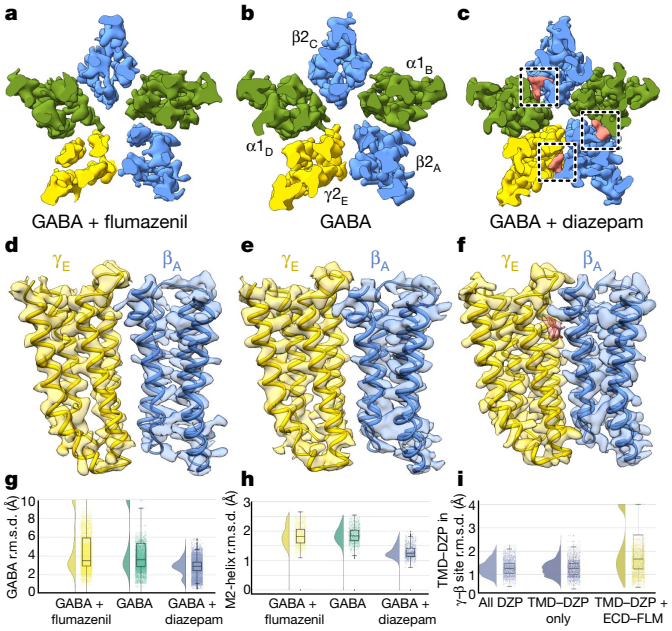


Fig. 3 | Binding sites of benzodiazepines and their mechanism of action. **a**–**c**, *Z*-slices in the TMD of cryo-EM density maps for the receptor in complex with GABA plus flumazenil (**a**), GABA alone (**b**) and GABA plus diazepam (**c**). Boxes in **c** highlight diazepam (salmon) TMD sites. **d**–**f**, Map and model of the TMD at the γ–β interface for the respective complexes in **a**–**c**, illustrating the large interfacial gap in the complex with GABA plus flumazenil, the smaller gap in the complex with GABA alone, and the absence of a gap in the complex with GABA plus diazepam. **g**–**i**, The stability of the stated components in benzodiazepine-related simulations, with probability distribution on the left, and raw data (n = 500 samples from 4 simulations, see Methods) with box plots indicating the median, interquartile range (25th–75th percentiles) and minimum–maximum ranges on the right. **g**, Simulations with bound diazepam (blue) exhibit stabilization of GABA over both orthosteric sites relative to the flumazenil-bound (yellow) or GABA-alone (green) conditions. **h**, Stabilization of M2 helices in the presence of diazepam. **i**, Destabilization of the transmembrane γ–β interface in the presence of extracellular flumazenil, relative to either diazepam or no ligand at the extracellular α–γ interface. DZP, diazepam; FLM, flumazenil.

ECD (Supplementary Video 5). In simulations of the complex with GABA plus flumazenil and the complex with GABA alone, GABA frequently dissociated (Fig. 3g, Extended Data Fig. 7h); conversely, GABA remained stably bound at both its binding sites in all simulations of the complex with GABA plus diazepam (Fig. 3g, Extended Data Fig. 7i). Diazepam also stabilized the TMD, as assessed by the root-mean-square deviation (r.m.s.d.) of the pore-lining M2 helices relative to that of complexes with GABA alone and with GABA plus flumazenil (Fig. 3h). We next simulated the substitution of flumazenil for diazepam at the ECD site while preserving the TMD diazepam molecules, and observed specific dissociation of diazepam from the γ–β site (Fig. 3i), consistent with our structure-based hypothesis that flumazenil binding in the ECD destabilizes this interface. Taken together, structural and dynamic analyses reveal that both benzodiazepine and anaesthetic positive modulators stabilize local and global organization of the receptor.

Comparison with recent structures

The structure of the α1β2γ2 receptor in complex with GABA plus diazepam provides an opportunity for direct comparison with that of the α1β3γ2 receptor in complex with the same ligands[19]. There are important differences in the approaches used to obtain these structures, including the use of a truncation in the M3–M4 loop in the constructs of our studies (discussed in Extended Data Figs. 8–10 and Supplementary Information). Sequence identity between β2 and β3 is 92% when the mostly disordered intracellular domain is not considered. Consistencies lend confidence to the results and differences might have important consequences for physiology or model interpretation. Overall, the functional profiles (Extended Data Fig. 11) and the structures agree well in architectural details—including pose and interactions of the ligands (Extended Data Fig. 8a, b)—except as noted in the distinct diazepam-binding site in the TMD. Global comparisons show that the TMD of the α1β3γ2 receptor is more compact and its pore is more constricted (Extended Data Fig. 9a–c). We sought additional reference points for direct comparison and obtained cryo-EM structures of the α1β2γ2 receptor in complex with the competitive antagonist bicuculline (methylated form) at 3.1 Å resolution, and with GABA plus the channel blocker picrotoxin at 2.9 Å resolution (Extended Data Figs. 3, 4, 6, 8, Extended Data Tables 1, 2, Supplementary Information). We observed the same trend in pore constriction in these structures as for the α1β3γ2 structures, although the top of the pore in our structures is consistently wider (Extended Data Fig. 9d–i).

Extending the comparison to other members of the Cys-loop receptor superfamily, the structures of the α1β3γ2 receptor have more surface area buried at subunit interfaces than any other structures in the anion-selective branch (Extended Data Fig. 9j). Examination of low-pass-filtered maps reveals a smaller nanodisc diameter for the α1β3γ2 structures (90–93 Å) compared with the α1β2γ2 structures (107–109 Å; Extended Data Fig. 10a–c). This finding was surprising because the scaffold used for the former structures, MSP2N2, was used intentionally for its large diameter[42] of around 150–165 Å; however, in the α1β3γ2 maps, it wraps tightly around the TMD. Differences in reconstitution may underlie the discrepancy: the on-column reconstitution approach used in the α1β3γ2 studies[18,19] removes excess lipids, while still in detergent, before adding the nanodisc scaffold. As detergent is removed, the scaffold could condense around the TMD. By contrast, in an effort to better mimic a physiological membrane, we included excess lipids throughout purification and reconstitution (Methods, Extended Data Fig. 10d). The result is a layer of lipids that insulate the α1β2γ2 receptor from the saposin shell, and more flexibility and a wider pore in the TMD. The differences are relatively subtle but are systematic, and could help to explain why the pore conformation of the α1β3γ2 structures in the presence of picrotoxin with or without GABA, and in the presence of bicuculline, are essentially identical[19]—unlike the two distinct conformations we observe (Supplementary Figs. 2–4, Supplementary Discussion). We suggest that delipidation during reconstitution of the α1β3γ2 $GABA_A$ receptor constrained the TMD and obscured the full range of conformational changes.

Conformational state and anaesthetic selectivity

The pores of all six structures involving agonists and agonists plus modulators adopt desensitized conformations, with a closed gate at the base of the pore at the level of the −2′ side chains (Extended Data Fig. 8f, g), consistent with expectations from steady-state physiological responses. Notably, all structures featuring bound intravenous anaesthetics show an increase in channel diameter at the 9′ position relative to those with GABA alone. This expansion of the pore at its midpoint results from rotation of the 9′ leucine sidechains away from the central axis, towards the adjacent subunit, leading to a decrease in the free-energy barrier to chloride permeation (Extended Data Fig. 8h). This rotation is a hallmark of activation[43], suggesting that the potentiation mechanism of intravenous anaesthetics could include stabilizing the 9′ activation gate in an open-like state (Extended Data Fig. 8g).

The pore conformations in the presence of GABA plus picrotoxin, and in the presence of the competitive antagonist bicuculline, contrast with these desensitized states. Bicuculline stabilizes a closed, resting-like state of the pore with a closed gate at the 9′ position (Extended Data Fig. 8g), similar to that observed in the α1β3γ2 structure[19]; relative to

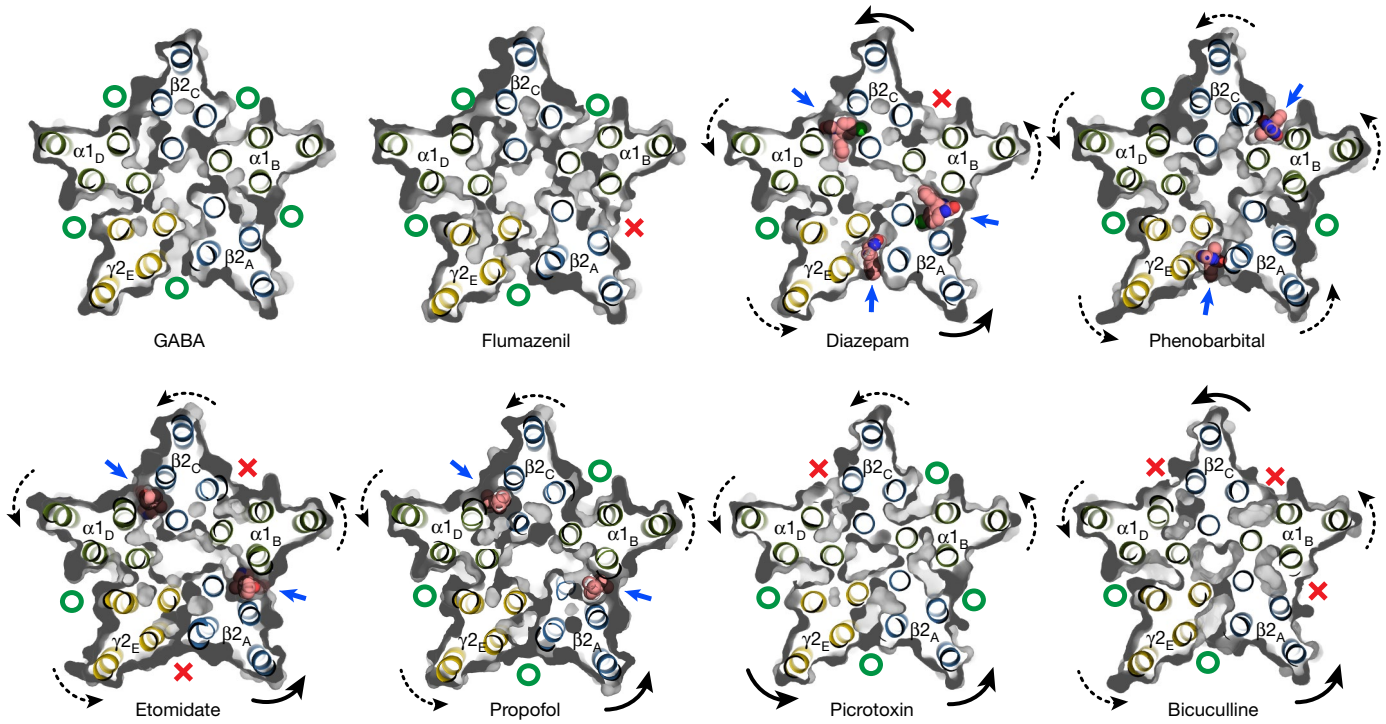


Fig. 4 | Selectivity and conformation of the anaesthetic cavity. Models and molecular surfaces are shown from a perspective down the channel axis, at the level of the TMD binding sites identified for diazepam, phenobarbital, etomidate and propofol. Straight arrows indicate occupied binding sites with ligands shown as spheres. Curved arrows indicate rigid-body subunit transformations relative to the structure of the complex with GABA alone (dashed lines are minor; solid lines are major rotations or translations). Green open circles indicate open or partially open cavities; red crosses indicate closed-off cavities.

GABA complexes, this structure is less hydrated above the hydrophobic gate in molecular dynamics simulations (Supplementary Fig. 2a). Picrotoxin, in the presence of GABA, stabilizes what we suggest is an intermediate state between the desensitized and resting states, in which the ECD adopts a compact agonist-bound conformation while the TMD adopts a more resting-like conformation in which the 9′ gate is partially closed. The results of electrophysiology experiments are consistent with this structural interpretation, as are comparisons of buried surface areas at interfaces (Extended Data Fig. 9k, Supplementary Fig. 3) and published observations that picrotoxin readily dissociates from the agonist-bound receptor[44]. Simulations of the picrotoxin-bound structure also demonstrated a similar extent of hydration above the hydrophobic gate to that of other GABA complexes, greater than in the bicuculline complex (Supplementary Fig. 2a). Furthermore, principal component analysis of TMD transitions project the picrotoxin-bound structure along a path from GABA-bound to bicuculline-bound states (Extended Data Fig. 8i, Supplementary Fig. 2b); principal component analysis within the ECD clustered the picrotoxin complex with GABA alone (Supplementary Fig. 2c). Thus, structural, functional and simulation results are consistent with picrotoxin—in the presence of GABA—binding to a receptor that has a desensitized- or activated-like ECD conformation and an intermediate TMD conformation, from which it can dissociate more readily than it could from a simple resting state.

Although the funnel shape of the TMD pore is similar among the GABA-bound and modulator-bound structures, we observed modulator-induced asymmetric motions that correlate with the specific site(s) occupied by a specific ligand. These subunit transformations are complex and include translations and rotations, with all rotations anticlockwise about an axis approximately normal to the membrane plane and through varying positions of each subunit. The transformations result in the opening or closing of access to the anaesthetic TMD pockets (Fig. 4; curved arrows approximate the trend in transformation). In the structure in which GABA alone is bound, all five interfacial sites are open. The binding of flumazenil has little overall effect, but by destabilizing the γ–β interface it leads to closure of the β–α TMD pocket in which diazepam, etomidate and propofol bind. This observation suggests a compelling long-range allosteric mechanism for anaesthetic reversal upon clinical administration of flumazenil[45,46]. Diazepam—through binding at one ECD site and three TMD sites—promotes global rotation of the TMD halves of all subunits, most noticeably in the β2 subunits, which results in closure of the α–β interface pocket. Phenobarbital causes a less marked but more symmetric rotation of all subunits through binding at the γ–β and α–β interfaces. Etomidate and propofol bind at common β–α sites; etomidate closes both γ–β and α–β access points whereas binding of the smaller propofol does not affect access to other sites. Notably, all potentiator-bound structures (phenobarbital, etomidate, propofol and diazepam) clustered in a region along the dominant principal components of motion for the TMD that was distinct from that of complexes with inhibitors (bicuculline or picrotoxin), flumazenil, or with GABA alone (Extended Data Fig. 8i, Supplementary Fig. 2b).

Picrotoxin binding results in occlusion of a single β–α interface, whereas bicuculline binding closes both β–α interfaces as well as the α–β site, further emphasizing the distinction in conformational states between the picrotoxin-bound and bicuculline-bound complexes. Bicuculline, in addition to inducing a rotation in the TMD halves of the subunits, promotes a compression of the TMD that brings the 9′ leucine side chains into position to block ion permeation, and creates the most compact TMD structure among all of the structures (Extended Data Figs. 8g, 9k). The binding of bicuculline allosterically closes three of the anaesthetic pockets, including the α–β interface, which is consistent with its ability to partially antagonize receptor activation by

phenobarbital[47]. A notable observation is that one pocket, at the α–γ interface, is always open. We observe density consistent with a lipid at this position (Extended Data Fig. 5a, b, f, g), which may relate to why this pocket cannot be closed, and could explain why, at least among the panel of ligands we surveyed, none bind there. A speculative hypothesis is that a lipid plays the part of an endogenous modulator or cofactor at this site.

Taken together, this panel of structures illustrates the complex interplay between the binding of diverse modulators and conformational transitions. The observation of distinct, asymmetric structural differences arising from the binding of each ligand mirrors results from cysteine-accessibility, disulfide crosslinking and electrophysiology studies that uncovered functional asymmetry in structural transitions[22,33,48,49]. The structural and dynamic stabilization that results from the binding of anaesthetics and benzodiazepines as positive modulators contrasts with the destabilization, in particular at the γ–β interface, observed upon the binding of flumazenil. The finding that flumazenil binding destabilizes the TMD of the receptor suggests a long-range allosteric mechanism for the reversal of the effects of benzodiazepines and anaesthetics by flumazenil.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2654-5.

1. Hemmings, H. C., Jr et al. Emerging molecular mechanisms of general anesthetic action. *Trends Pharmacol. Sci.* **26**, 503–510 (2005).
2. Forman, S. A. & Miller, K. W. Mapping general anesthetic sites in heteromeric γ-aminobutyric acid type A receptors reveals a potential for targeting receptor subtypes. *Anesth. Analg.* **123**, 1263–1273 (2016).
3. Sieghart, W. & Savić, M. M. International Union of Basic and Clinical Pharmacology. CVI: $GABA_A$ receptor subtype- and function-selective ligands: key issues in translation to humans. *Pharmacol. Rev.* **70**, 836–878 (2018).
4. Sigel, E. & Ernst, M. The benzodiazepine binding sites of $GABA_A$ receptors. *Trends Pharmacol. Sci.* **39**, 659–671 (2018).
5. Olsen, R. W. $GABA_A$ receptor: positive and negative allosteric modulators. *Neuropharmacology* **136**, 10–22 (2018).
6. Meyer, H. Welche Eigenschaft der Anasthetica bedingt ihre narkotische Wirkung? *Naunyn Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **42**, 109–118 (1899).
7. Meyer, H. Zur Theorie der Alkoholnarkose: der Einfluss wechselnder Temperatur auf Wirkungsstärke und Theilungscoefficient der Narcotica. *Naunyn Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **46**, 338–346 (1901).
8. Overton, E. *Studien über die Narkose Zugleich ein Beitrag zur allgemeinen Pharmakologie* (Gustav Fischer, 1901).
9. Janoff, A. S., Pringle, M. J. & Miller, K. W. Correlation of general anesthetic potency with solubility in membranes. *Biochim. Biophys. Acta* **649**, 125–128 (1981).
10. Franks, N. P. & Lieb, W. R. Molecular and cellular mechanisms of general anaesthesia. *Nature* **367**, 607–614 (1994).
11. Krasowski, M. D. & Harrison, N. L. General anaesthetic actions on ligand-gated ion channels. *Cell. Mol. Life Sci.* **55**, 1278–1303 (1999).
12. Krasowski, M. D. Contradicting a unitary theory of general anesthetic action: a history of three compounds from 1901 to 2001. *Bull. Anesth. Hist.* **21**, 1–24 (2003).
13. Mihic, S. J. et al. Sites of alcohol and volatile anaesthetic action on $GABA_A$ and glycine receptors. *Nature* **389**, 385–389 (1997).
14. Drexler, B., Antkowiak, B., Engin, E. & Rudolph, U. Identification and characterization of anesthetic targets by mouse molecular genetics approaches. *Can. J. Anaesth.* **58**, 178–190 (2011).
15. Walters, R. J., Hadley, S. H., Morris, K. D. & Amin, J. Benzodiazepines act on $GABA_A$ receptors via two distinct and separable mechanisms. *Nat. Neurosci.* **3**, 1274–1281 (2000).
16. Middendorp, S. J., Maldifassi, M. C., Baur, R. & Sigel, E. Positive modulation of synaptic and extrasynaptic $GABA_A$ receptors by an antagonist of the high affinity benzodiazepine binding site. *Neuropharmacology* **95**, 459–467 (2015).
17. Votey, S. R., Bosse, G. M., Bayer, M. J. & Hoffman, J. R. Flumazenil: a new benzodiazepine antagonist. *Ann. Emerg. Med.* **20**, 181–188 (1991).
18. Laverty, D. et al. Cryo-EM structure of the human α1β3γ2 $GABA_A$ receptor in a lipid bilayer. *Nature* **565**, 516–520 (2019).
19. Masiulis, S. et al. $GABA_A$ receptor signalling mechanisms revealed by structural pharmacology. *Nature* **565**, 454–459 (2019).
20. Löscher, W. & Rogawski, M. A. How theories evolved concerning the mechanism of action of barbiturates. *Epilepsia* **53**, 12–25 (2012).
21. Zhu, S. et al. Structure of a human synaptic $GABA_A$ receptor. *Nature* **559**, 67–72 (2018).
22. Gielen, M., Barilone, N. & Corringer, P.-J. The desensitization pathway of $GABA_A$ receptors, one subunit at a time. Preprint at https://doi.org/10.1101/2020.05.18.101287 (2020).
23. Chiara, D. C. et al. Specificity of intersubunit general anesthetic-binding sites in the transmembrane domain of the human α1β3γ2 γ-aminobutyric acid type A ($GABA_A$) receptor. *J. Biol. Chem.* **288**, 19343–19357 (2013).
24. Zeller, A., Arras, M., Jurd, R. & Rudolph, U. Identification of a molecular target mediating the general anesthetic actions of pentobarbital. *Mol. Pharmacol.* **71**, 852–859 (2007).
25. Belelli, D., Callachan, H., Hill-Venning, C., Peters, J. A. & Lambert, J. J. Interaction of positive allosteric modulators with human and *Drosophila* recombinant GABA receptors expressed in *Xenopus laevis* oocytes. *Br. J. Pharmacol.* **118**, 563–576 (1996).
26. Vuyk, J., Sitsen, E. & Reekers, M. in *Miller's Anesthesia* 9th edn (eds Gropper, M. A. et al.) Ch. 23, 638–679 (Elsevier, 2020).
27. Forman, S. A. Clinical and molecular pharmacology of etomidate. *Anesthesiology* **114**, 695–707 (2011).
28. Belelli, D., Lambert, J. J., Peters, J. A., Wafford, K. & Whiting, P. J. The interaction of the general anesthetic etomidate with the γ-aminobutyric acid type A receptor is influenced by a single amino acid. *Proc. Natl Acad. Sci. USA* **94**, 11031–11036 (1997).
29. Siegwart, R., Jurd, R. & Rudolph, U. Molecular determinants for the action of general anesthetics at recombinant $\alpha_2\beta_3\gamma_2$γ-aminobutyric $acid_A$ receptors. *J. Neurochem.* **80**, 140–148 (2002).
30. Li, G. D. et al. Identification of a $GABA_A$ receptor anesthetic binding site at subunit interfaces by photolabeling with an etomidate analog. *J. Neurosci.* **26**, 11599–11605 (2006).
31. Krasowski, M. D. et al. Propofol and other intravenous anesthetics have sites of action on the γ-aminobutyric acid type A receptor distinct from that for isoflurane. *Mol. Pharmacol.* **53**, 530–538 (1998).
32. Jayakar, S. S. et al. Multiple propofol-binding sites in a γ-aminobutyric acid type A receptor ($GABA_AR$) identified using a photoreactive propofol analog. *J. Biol. Chem.* **289**, 27456–27468 (2014).
33. Bali, M. & Akabas, M. H. Gating-induced conformational rearrangement of the γ-aminobutyric acid type A receptor β-α subunit interface in the membrane-spanning domain. *J. Biol. Chem.* **287**, 27762–27770 (2012).
34. Jayakar, S. S. et al. Identifying drugs that bind selectively to intersubunit general anesthetic sites in the α1β3γ2 $GABA_AR$ transmembrane domain. *Mol. Pharmacol.* **95**, 615–628 (2019).
35. Yip, G. M. et al. A propofol binding site on mammalian $GABA_A$ receptors identified by photolabeling. *Nat. Chem. Biol.* **9**, 715–720 (2013).
36. Jurd, R. et al. General anesthetic actions in vivo strongly attenuated by a point mutation in the $GABA_A$ receptor β3 subunit. *FASEB J.* **17**, 250–252 (2003).
37. Reynolds, D. S. et al. Sedation and anesthesia mediated by distinct $GABA_A$ receptor isoforms. *J. Neurosci.* **23**, 8608–8617 (2003).
38. Chiara, D. C. et al. Mapping general anesthetic binding site(s) in human α1β3 γ-aminobutyric acid type A receptors with [^{3}H]TDBzl-etomidate, a photoreactive etomidate analogue. *Biochemistry* **51**, 836–847 (2012).
39. Krasowski, M. D., Hong, X., Hopfinger, A. J. & Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the $GABA_A$ receptor. *J. Med. Chem.* **45**, 3210–3221 (2002).
40. Krasowski, M. D., Nishikawa, K., Nikolaeva, N., Lin, A. & Harrison, N. L. Methionine 286 in transmembrane domain 3 of the $GABA_A$ receptor beta subunit controls a binding cavity for propofol and other alkylphenol general anesthetics. *Neuropharmacology* **41**, 952–964 (2001).
41. Eaton, M. M. et al. Multiple non-equivalent interfaces mediate direct activation of $GABA_A$ receptors by propofol. *Curr. Neuropharmacol.* **14**, 772–780 (2016).
42. Ritchie, T. K. et al. Chapter eleven - reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods Enzymol.* **464**, 211–231 (2009).
43. Damgen, M. A. & Biggin, P. C. A refined open state of the glycine receptor obtained via molecular dynamics simulations. *Structure* **28**, 130–139.e2 (2020).
44. Gielen, M., Thomas, P. & Smart, T. G. The desensitization gate of inhibitory Cys-loop receptors. *Nat. Commun.* **6**, 6829 (2015).
45. Dahaba, A. A. et al. Effect of flumazenil on bispectral index monitoring in unpremedicated patients. *Anesthesiology* **110**, 1036–1040 (2009).
46. Safavynia, S. A. et al. Effects of γ-aminobutyric acid type A receptor modulation by flumazenil on emergence from general anesthesia. *Anesthesiology* **125**, 147–158 (2016).
47. Ueno, S., Bracamontes, J., Zorumski, C., Weiss, D. S. & Steinbach, J. H. Bicuculline and gabazine are allosteric inhibitors of channel opening of the $GABA_A$ receptor. *J. Neurosci.* **17**, 625–634 (1997).
48. Baumann, S. W., Baur, R. & Sigel, E. Individual properties of the two functional agonist sites in $GABA_A$ receptors. *J. Neurosci.* **23**, 11158–11166 (2003).
49. Rosen, A., Bali, M., Horenstein, J. & Akabas, M. H. Channel opening by anesthetics and GABA induces similar changes in the $GABA_A$ receptor M2 segment. *Biophys. J.* **92**, 3130–3139 (2007).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Receptor expression and purification

A tri-cistronic construct of the human α1β2γ2 $GABA_A$ receptor, with the three genes linked by a 22-amino-acid-long P2A 'self-cleaving' peptide[50], was designed, codon-optimized, synthesized and cloned into the pEZT-BM expression vector to enhance the expression of the tri-heteromeric receptor[51]. Both a full-length wild-type and M3–M4-loop-truncation construct were made in this tri-cistronic format. In the construct used for EM, the M3–M4 loop of each subunit was replaced by a linker peptide, SQPARAA, as in our previous study[21]. The order of the subunits in the expression construct was β2-γ2-α1, with a twin Strep tag placed at the N terminus of the γ2 subunit for purification. BacMam virus was produced using Sf9 cells (ATCC CRL-1711) and titred as described for the α4β2 nicotinic receptor[51]. Suspension cultures of HEK293S GnTI$^-$ cells (ATCC CRL-3022) were grown at 37 °C with 8% CO_2 and were transduced with multiplicities of infection of 0.5 at a cell density of 3.5×10^6–4.0×10^6 cells per ml. The HEK and Sf9 cell lines were not authenticated nor were they tested for mycoplasma. At the time of transduction, 1 mM sodium butyrate (Sigma-Aldrich) was added to the culture and the temperature was reduced to 30 °C to enhance protein expression. After 72 h, cells were collected by centrifugation and resuspended in 20 mM Tris, pH 7.4, 150 mM NaCl (TBS buffer) containing 1 mM phenylmethanesulfonyl fluoride (PMSF; Sigma-Aldrich) and the target ligands (2 mM GABA; 1 μM flumazenil (Santa Cruz Biotechnology) + 2 mM GABA; 200 μM diazepam (Sigma-Aldrich) + 2 mM GABA; 2 mM phenobarbital (Sigma-Aldrich) + 2 mM GABA; 500 μM etomidate (Tocris) + 2 mM GABA; 100 μM propofol (Sigma-Aldrich) + 2 mM GABA; 50 μM bicuculline methbromide (Sigma-Aldrich); 100 μM picrotoxin (Sigma-Aldrich) + 2 mM GABA) for the intended complex, and lysed using an Avestin Emulsiflex. Lysed cells were centrifuged for 20 min at 10,000*g* and the resulting supernatants were centrifuged at 186,000*g* for 2 h. Membrane pellets were homogenized using a Dounce homogenizer and solubilized in TBS buffer containing 40 mM *n*-dodecyl-β-maltoside (DDM, Anatrace) and 1 mM PMSF and ligands. Solubilized membranes were centrifuged for 40 min at 186,000*g* and the supernatants were passed over Strep-Tactin XT Superflow affinity resin (IBA-GmbH). The resin was washed with TBS buffer containing 0.01% (w/v) porcine brain polar lipids (Avanti), 2 mM DDM and ligands. The receptors were eluted in the same buffer containing 50 mM biotin (Sigma-Aldrich).

Receptor–nanodisc reconstitution

The saposin A expression plasmid was provided by Salipro Biotech AB. We selected saposin over other nanodisc scaffolds because of its ability to accommodate a range of membrane protein sizes and preserve an approximately symmetric TMD conformation (Extended Data Fig. 1a–e). Reconstitution of $GABA_A$ receptors into saposin-based nanodiscs was modified from a previously published protocol[52] (Extended Data Fig. 10d). The concentrated α1β2γ2 receptors (~15 μM) were pre-incubated with porcine brain polar lipids for 10 min at room temperature, and then saposin was added and incubated for 2 min. The molar ratio of receptor, lipids and saposin was 1:230:30. The reaction was diluted approximately 10-fold by TBS to initiate reconstitution. Detergent was removed by adding Bio-Beads SM-2 (Bio-Rad) to a final concentration of 200 mg ml^{-1} while rotating overnight at 4 °C. Bio-Beads were removed the next day, and the sample was collected for size-exclusion chromatography.

Cryo-EM sample preparation

The α1β2γ2 receptors reconstituted in nanodiscs were mixed with 1F4 Fab[21] in a 3:1 (w/w) ratio. After incubating for 15 min, the mixture was concentrated and injected over a Superose 6 Increase 10/300 GL column (GE Healthcare) equilibrated in TBS with ligands (2 mM GABA; 1 μM flumazenil + 2 mM GABA; 200 μM diazepam + 2 mM GABA; 2 mM phenobarbital + 2 mM GABA; 500 μM etomidate + 2 mM GABA; 100 μM propofol + 2 mM GABA; 100 μM picrotoxin + 2 mM GABA; 50 μM bicuculline methobromide). Peak fractions were analysed by fluorescence-detection size-exclusion chromatography, monitoring the fluorescence of tryptophan. Fractions showing a single peak were collected and concentrated to an absorbance at 280 nm (A_{280}) of 7–9. During sample concentration, the buffer for the propofol sample was changed to TBS with 2 mM GABA and 1 mM propofol. Immediately before freezing grids, 0.5 mM fluorinated Fos-Choline-8 (Anatrace) was mixed with the sample to minimize preferred orientation. Then, 3 μl of sample was applied to glow-discharged gold R1.2/1.3 200 mesh holey carbon grids (Quantifoil) and immediately blotted for 3 s at 100% humidity and 4 °C. The grids were then plunge-frozen into liquid ethane using a Vitrobot Mark IV (FEI).

Cryo-EM data collection and processing

Cryo-EM data were collected on a 300 kV Titan Krios Microscope (FEI) equipped with a K2 Summit or a K3 direct electron detector (Gatan) and a GIF quantum energy filter (20 eV) (Gatan) using super-resolution mode. Details of all datasets are summarized in Extended Data Tables 1 and 2. All datasets were processed using the same general workflow in RELION 3.0 or 3.1[53]. Dose-fractionated images were gain normalized, 2× Fourier binned, aligned, dose-weighted and summed using MotionCor2[54]. Contrast transfer function (CTF) and defocus value estimation were performed using GCTF[55] or CTFFIND4[56]. Particle picking for the three datasets collected at the Harvard Medical School (HMS) facility carried out using crYOLO[57]. For the five datasets collected at the University of Texas Southwestern (UTSW) and the Pacific Northwest Center for Cryo-EM (PNCC) facilities, around 50 particles were picked manually and subjected to reference-free 2D classification to generate initial references for autopicking in RELION. These references were then used for autopicking from a subset of 30–50 images, and then 2D classification was repeated to obtain good references for autopicking on all images. After autopicking, images were inspected, and bad images and false-positive particles were removed manually and by particle sorting. Ab initio models were generated using 3,000–5,000 good particles in RELION, and then were used for 3D classification. 3D classes with strong TMD signal were selected for 3D refinement. The best 3D class was used for an initial model (low-pass-filtered to 40 or 50 Å) for 3D refinement. Per-particle CTF refinement and beam tilt estimation were performed and a second round of refinement was followed by fine local angular sampling using the map from the first refinement as the initial model, which was low-pass-filtered to 10 Å. Because we observed a high level of disorder in the TMD of the γ-subunit in all eight datasets, focused 3D classification without alignment[58] was performed on the γ-TMD after subtracting the signal from the rest of the receptor and nanodisc. Particles from the best classes were selected for particle polishing and an additional round of 3D refinement to generate the final maps. Local resolution was estimated with ResMap[59].

Model building, refinement and validation

An initial model was generated by combining the ECD of the heteropentameric $GABA_A$ receptor–Fab complex bound to GABA + flumazenil (RCSB: 6D6U)[21] and the TMD of a homology model generated by Swiss-Model[60] based on the β3 homopentamer structure (RCSB: 4COF)[61]. This model was docked into the density map using UCSF Chimera[62]. The model was manually adjusted, and flumazenil was removed, in Coot[63]. To build models of the different complexes, the GABA-bound structure was first built and then used as a starting model. Well-ordered N-linked glycans were built within the vestibule and along the surface of the ECD. In GABA, GABA + diazepam, GABA + etomidate, GABA + propofol and bicuculline complexes, an additional branch of mannose

densities was found in chain B and built de novo. After manual building in Coot, global real space and B-factor refinement with stereochemistry restraints were performed in Phenix[64]. The map-model FSC value between the final model and the map was estimated by Phenix and plotted in Extended Data Fig. 3. In Extended Data Tables 1 and 2, we list the fraction of particles used in the final fraction relative to those that emerge from 2D classification. We found a correlation between this percentage and relative order in the γ2-TMD. The GABA + flumazenil reconstruction was produced from only 8% of the particles selected after 2D classification, indicating a high degree of intrinsic flexibility in the γ-TMD. In the GABA-alone complex, the fraction was 16%. For the diazepam complex, 36% of the particles after 2D classification had a well-ordered γ2-TMD, similar to that from the phenobarbital complex, providing a measure of the increase in stability in the diazepam complex relative to both flumazenil and GABA alone.

Schematic interaction analysis of the bound ligands was performed using Ligplot+[65]. Subunit interfaces were analysed by PDBePISA server[66]. Pore radius profiles were analysed using Hole2[67]. Sequence alignments were made using PROMALS3D[68]. Structural figures were generated by UCSF Chimera and PyMOL (Schrodinger, LLC). Structural biology software packages were compiled by SBGrid[69].

Electrophysiology

Whole-cell voltage-clamp recordings were made from adherent HEK293S GnTI$^-$ cells transiently transfected with the tri-cistronic pEZT construct used for structural analysis. Upon transfection with 0.2–0.5 μg of the plasmid per well in a 12-well dish, the cells were transferred to 30 °C. On the day of recording (1–3 days later), cells were re-plated onto a 35 mm dish and washed with bath solution, which contained (in mM): 140 NaCl, 2.4 KCl, 4 $MgCl_2$, 4 $CaCl_2$, 10 HEPES pH 7.3 and 10 glucose. Borosilicate pipettes were pulled and polished to an initial resistance of 2–4 MΩ. The pipette solution contained (in mM): 150 CsCl, 10 NaCl, 10 EGTA and 20 HEPES pH 7.3. Cells were clamped at −75 mV. The recordings were made with an Axopatch 200B amplifier, sampled at 5 kHz, and low-pass-filtered at 2 kHz using a Digidata 1440A (Molecular Devices) and analysed with pClamp 10 software (Molecular Devices). The ligand solutions were prepared in bath solution from concentrated stocks. Stocks of 1 M GABA and 500 mM phenobarbital were prepared in water and 100 mM stocks of bicuculline, diazepam, picrotoxin, etomidate and 10 mM stock of flumazenil were prepared in DMSO. Solution exchange was achieved using a gravity-driven RSC-200 rapid solution changer (Bio-Logic). In phenobarbital potentiation experiments with mutants (Fig. 1b), responses are from 5 μM GABA compared to 5 μM GABA plus 500 μM phenobarbital. Peak currents were measured using HEK293S GnTI$^-$ expressing wild-type or mutant receptors. The experiments were repeated at least 3 times from three different cells. Statistical analyses were performed using Prism v8 (GraphPad). To quantify differences in peak currents between EM and mutant constructs, mean and standard deviations were calculated from more than three independent patches for each group. An unpaired two-tailed Student's *t*-test was used for single comparisons between wild-type and mutant groups. * and ** denote statistical significance corresponding to *P* values of <0.01 and <0.0001, respectively.

Coarse-grained simulations

Atomic coordinates for the α1β2γ2 receptor with the intracellular domain modification in complex with GABA plus phenobarbital were coarse-grained, through the representation of about four heavy atoms as a single bead, using Martini Bilayer Maker in CHARMM-GUI[70]. Ligands and glycans were omitted, and the protein was embedded in a symmetric membrane containing 40% cholesterol, 20% 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC), 20% 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine (POPE), 9% 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-L-serine (POPS) and 1% phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P_2), previously shown to approximate the neuronal plasma membrane[24]. In total, 4,437 lipids were inserted in the simulation system, constituting 313,112 total beads including water and ions. After energy minimization and equilibration in CHARMM-GUI, simulations were run with the protein restrained for 25 μs in GROMACS 2019.4[71] to allow lipid convergence, using Martini 2.2 and 2.0 parameters[72] for amino acids and lipids, respectively. Five replicates were performed from different initial lipid compositions generated in CHARMM-GUI. For comparison, an additional simulation of the receptor in complex with bicuculline was performed. All simulations relaxed within 20 μs to equivalent patterns of lipid association around the receptor, including local enrichment of PtdIns(4,5)P_2 and cholesterol at transmembrane subunit interfaces.

Molecular dynamics simulations

The final frame from a randomly selected coarse-grained simulation was selected for backmapping to an all-atom system, including all PtdIns(4,5)P_2 molecules observed to bind persistently in more than two replicates. The lipid bilayer was backmapped into CHARMM36 topologies[73], then placed around each protein model reported in this work, and trimmed to a box size of 14 × 14 × 16 nm. The system was solvated and neutralized in NaCl (approximately 150 mM). All atomistic simulations were performed using GROMACS 2019.4 in the CHARMM36 forcefield[74]. Simulations included resolved GABA or modulatory ligands except as indicated. For flumazenil-substitution simulations, flumazenil was superimposed from the flumazenil-bound structure in place of extracellular diazepam in the diazepam-bound structure; for propofol-saturation simulations in Extended Data Fig. 1, propofol was superimposed at the α–γ, γ–β and α–β interfaces on the basis of pseudo-symmetric poses at β–α interfaces in the propofol-bound structure. Parameters for ligand molecules were generated with CGenFF in CHARMM-GUI[70], with additional optimization using quantum mechanics for ligands with high penalty scores[75]. Each system was energy-minimized and then relaxed with a constant number of particles, pressure and temperature for at least 60 ns, during which the position restraints on the protein were gradually released. All ligands were restrained during equilibration. For each equilibrated system, four replicates of 500-ns unrestrained simulations were then generated and frames analysed every 4 ns, for a total of 500 samples in each condition (four replicates × 125 frames). The temperature was kept at 300 K using a velocity-rescaling thermostat[76], Parrinello–Rahman pressure coupling[77] ensured constant pressure, the particle mesh Ewald algorithm[78] was used for long-range electrostatic interactions, and hydrogen-bond lengths were constrained using the LINCS algorithm[79]. Analyses were performed using VMD[80], MDAnalysis[81] and MDTraj[82]. Simulation properties were represented using raincloud plots (https://doi.org/10.12688/wellcomeopenres.15191.1), for example, Fig. 3g–i, Extended Data Fig. 1h and Supplementary Fig. 2a showing unmirrored probability distribution functions on the left, and jittered raw data with superimposed box plots indicating sample median, interquartile range (25th–75th percentiles), minimum–maximum range, and outliers on the right.

Principal component analysis

Protein models in complex with GABA, bicuculline, GABA + etomidate, GABA + phenobarbital, and GABA + propofol were r.m.s.d.-aligned using all Cα atoms, then used to calculate principal components of motion in Cartesian coordinate space for Cα atoms of the TMD (residues equivalent to β2-218 to 338) or ECD (β2-10 to 217) in all subunits. Subsequently, all protein models reported in this work (n = 8 independent structures) were projected onto the PC1-2 subspaces for the two domains. Elastic-network interpolations between the bicuculline and GABA complexes were performed using eBDIMS[83] with cutoff = 6, mode = 3, and 1 unbiased step, then projected onto the principal component subspaces.

Ion permeation calculations

The free energy along the pore axis for chloride was calculated using the accelerated weight histogram (AWH) method[84]. In brief, for each

equilibrated structure (complexes with GABA, bicuculline, GABA + phenobarbital or GABA + propofol), we applied one independent AWH bias and simulated for 50 ns each with 16 walkers sharing bias data and contributing to the same target distribution. Each bias acts on the centre-of-mass z-distance between one central chloride ion and the Cα of β-270, α-275 and γ-285 residues, with a sampling interval across more than 95% of the box length along the z axis to reach periodicity. To keep the solute close to the pore entrance, the coordinate radial distance was restrained to stay below 10 Å by adding a flat-bottom umbrella potential.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Atomic model coordinates for bicuculline methbromide, GABA + propofol, GABA + flumazenil, GABA + etomidate, GABA + phenobarbital, GABA + diazepam, GABA and GABA + picrotoxin-bound structures have been deposited in the Protein Data Bank with accession codes 6X3S, 6X3T, 6X3U, 6X3V, 6X3W, 6X3X, 6X3Z and 6X40, respectively. Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank with accession codes EMD-22031, EMD-22032, EMD-22033, EMD-22034, EMD-22035, EMD-22036, EMD-22037 and EMD-22038, respectively.

50. Kim, J. H. et al. High cleavage efficiency of a 2Å peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS ONE* **6**, e18556 (2011).
51. Morales-Perez, C. L., Noviello, C. M. & Hibbs, R. E. Manipulation of subunit stoichiometry in heteromeric membrane proteins. *Structure* **24**, 797–805 (2016).
52. Lyons, J. A., Bøggild, A., Nissen, P. & Frauenfeld, J. Chapter three - saposin–lipoprotein scaffolds for structure determination of membrane transporters. *Methods Enzymol*. **594**, 85–99 (2017).
53. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
54. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
55. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol*. **193**, 1–12 (2016).
56. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol*. **192**, 216–221 (2015).
57. Wagner, T. et al. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol*. **2**, 218 (2019).
58. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
59. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
60. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res*. **31**, 3381–3385 (2003).
61. Miller, P. S. & Aricescu, A. R. Crystal structure of a human $GABA_A$ receptor. *Nature* **512**, 270–275 (2014).
62. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
63. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
64. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
65. Laskowski, R. A. & Swindells, M. B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model*. **51**, 2778–2786 (2011).
66. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol*. **372**, 774–797 (2007).
67. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph*. **14**, 354–360 (1996).
68. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res*. **36**, 2295–2300 (2008).
69. Morin, A. et al. Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
70. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem*. **29**, 1859–1865 (2008).
71. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
72. Periole, X. & Marrink, S. J. The Martini coarse-grained force field. *Methods Mol. Biol*. **924**, 533–565 (2013).
73. Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J. & Tieleman, D. P. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput*. **10**, 676–690 (2014).
74. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput*. **8**, 3257–3273 (2012).
75. Vanommeslaeghe, K. et al. CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem*. **31**, 671–690 (2010).
76. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys*. **126**, 014101 (2007).
77. Parrinello, M. & Rahman, A. Crystal structure and pair potentials: a molecular-dynamics study. *Phys. Rev. Lett*. **45**, 1196–1199 (1980).
78. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys*. **103**, 8577–8593 (1995).
79. Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput*. **4**, 116–122 (2008).
80. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph*. **14**, 33–38 (1996).
81. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem*. **32**, 2319–2327 (2011).
82. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J*. **109**, 1528–1532 (2015).
83. Orellana, L., Yoluk, O., Carrillo, O., Orozco, M. & Lindahl, E. Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. *Nat. Commun*. **7**, 12575 (2016).
84. Lindahl, V., Gourdon, P., Andersson, M. & Hess, B. Permeability and ammonia selectivity in aquaporin TIP2;1: linking structure to function. *Sci. Rep*. **8**, 2995 (2018).
85. Phulera, S. et al. Cryo-EM structure of the benzodiazepine-sensitive α1β1γ2S tri-heteromeric $GABA_A$ receptor in complex with GABA. *eLife* **7**, e39383 (2018).
86. Miller, P. S. et al. Heteromeric $GABA_A$ receptor structures in positively-modulated active states. Preprint at https://doi.org/10.1101/338343 (2018).
87. Ingólfsson, H. I. et al. Computational lipidomics of the neuronal plasma membrane. *Biophys. J*. **113**, 2271–2280 (2017).
88. Nury, H. et al. X-ray structures of general anaesthetics bound to a pentameric ligand-gated ion channel. *Nature* **469**, 428–431 (2011).
89. Fourati, Z. et al. Structural basis for a bimodal allosteric mechanism of general anesthetic modulation in pentameric ligand-gated ion channels. *Cell Rep*. **23**, 993–1004 (2018).
90. Hibbs, R. E. & Gouaux, E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**, 54–60 (2011).

Acknowledgements We thank R. Cabuco and L. Baxter for baculovirus production, and all members of the Hibbs laboratory for discussion. Single-particle cryo-EM data were collected at the University of Texas Southwestern Medical Center Cryo-Electron Microscopy Facility, which is supported by the CPRIT Core Facility Support Award RP170644, at the Harvard Cryo-Electron Microscopy Center for Structural Biology, and at the Pacific Northwest Cryo-EM Center at Oregon Health & Science University, which is supported by NIH grant U24GM129547, accessed through EMSL (grid.436923.9) a DOE office of Science User Facility sponsored by the Office of Biological and Environmental Research. Computational resources were provided by the Swedish National Infrastructure for Computing. J.J.K. and S.Z. acknowledge support from the American Heart Association grants 20POST35200127 and 18POST34030412, respectively. This work was supported by Vetenskapsrådet VR and the Knut and Alice Wallenberg foundation to E.L. and by The Welch Foundation (I-1812) and grants from the NIH (DA037492, DA042072, and NS095899) to R.E.H.

Author contributions J.J.K. and R.E.H. conceived the project. J.J.K. performed the construct design, protein production, purification, EM sample preparation and structural analysis including the EM data processing. J.J.K., A.G. and R.E.H. built the atomic models. J.T. performed the mutagenesis and electrophysiology experiments. S.Z., C.M.N. and R.M.W. collected the EM data. Y.Z., R.J.H. and E.L. performed and analysed simulations. J.J.K., A.G. and R.E.H. wrote the manuscript with input from all other authors.

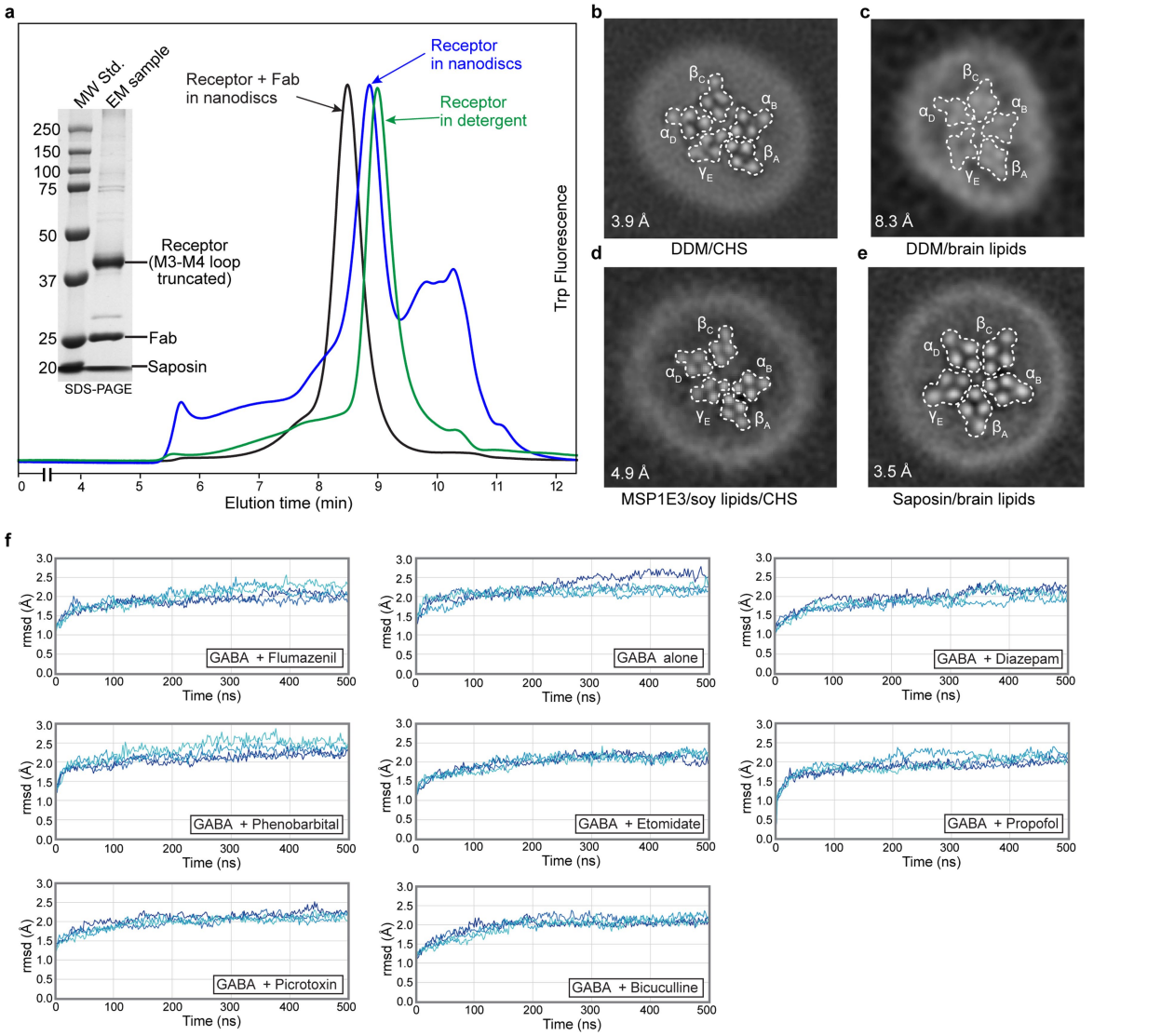
Competing interests The authors declare no competing interests.

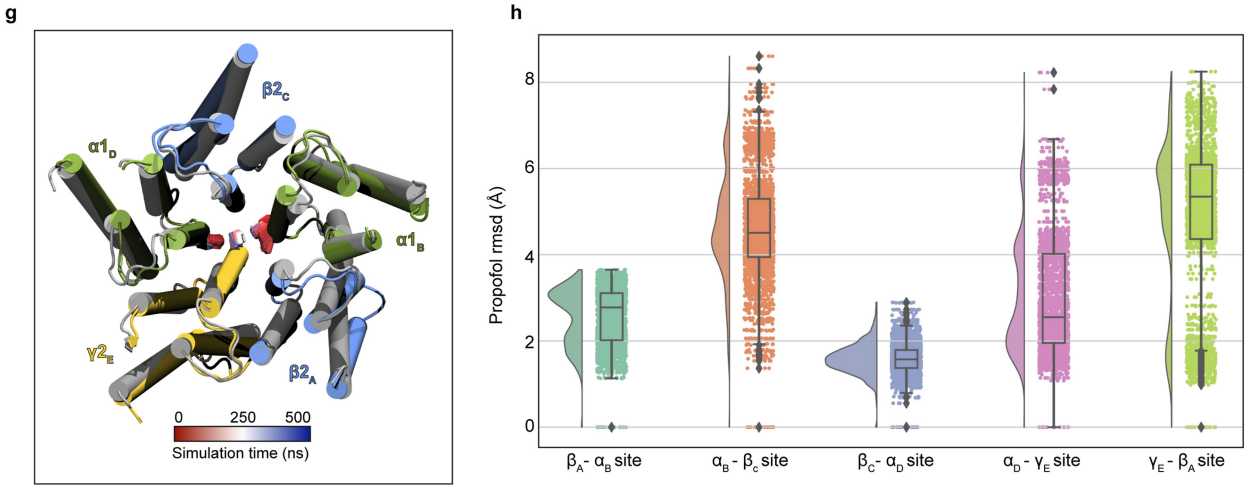
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2654-5.

Correspondence and requests for materials should be addressed to R.E.H.

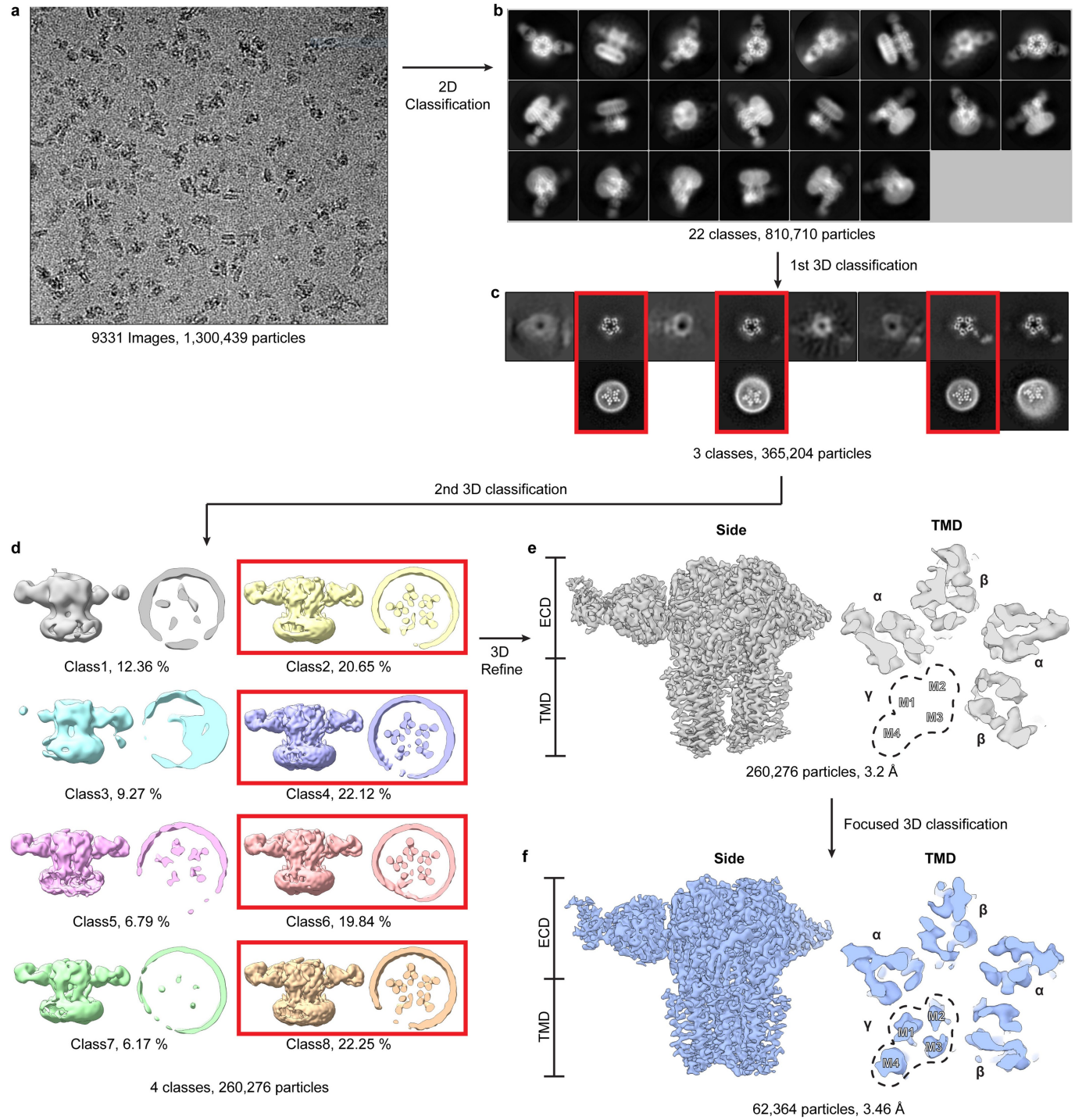
Peer review information *Nature* thanks Margot Ernst and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



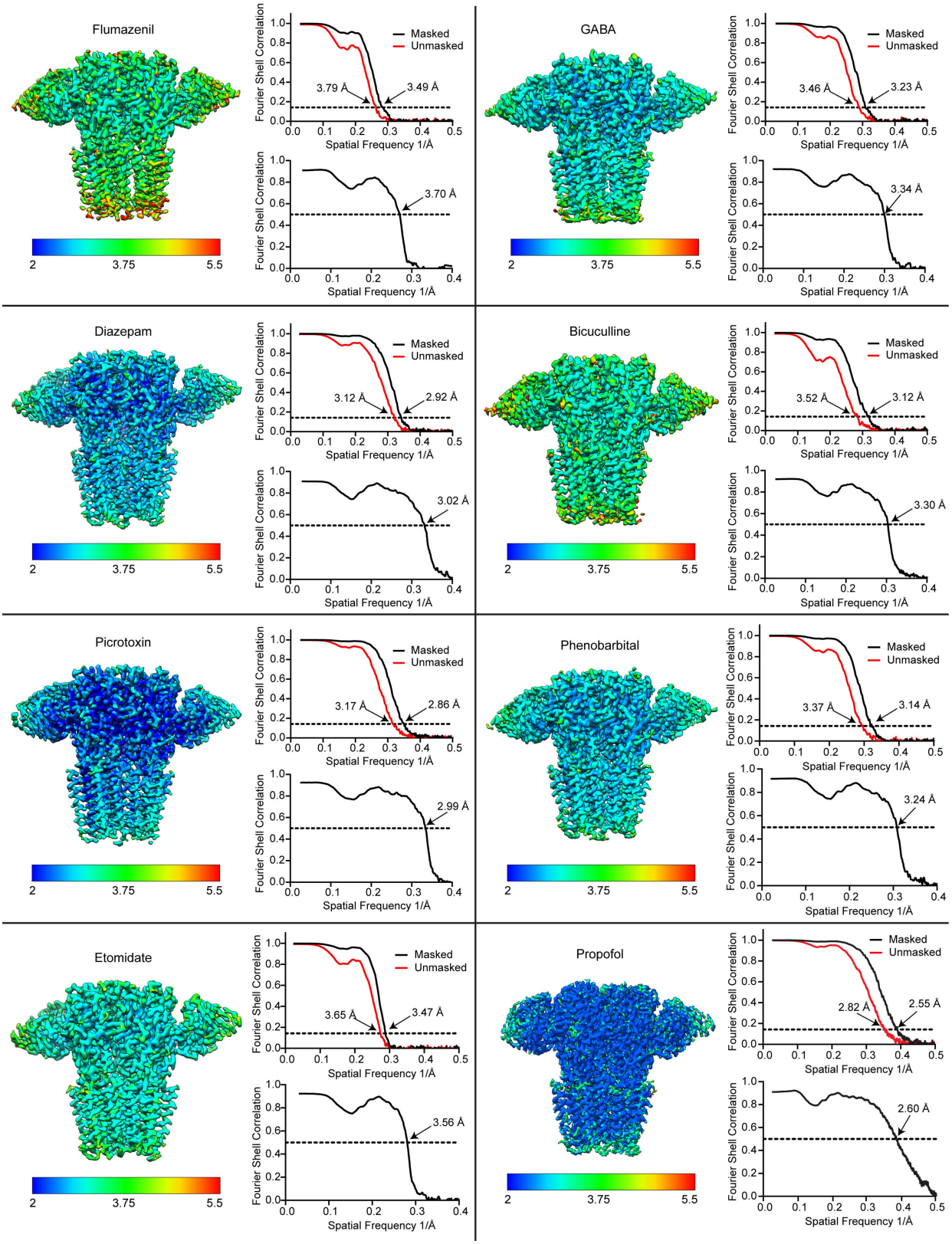


Extended Data Fig. 1 | See next page for caption.

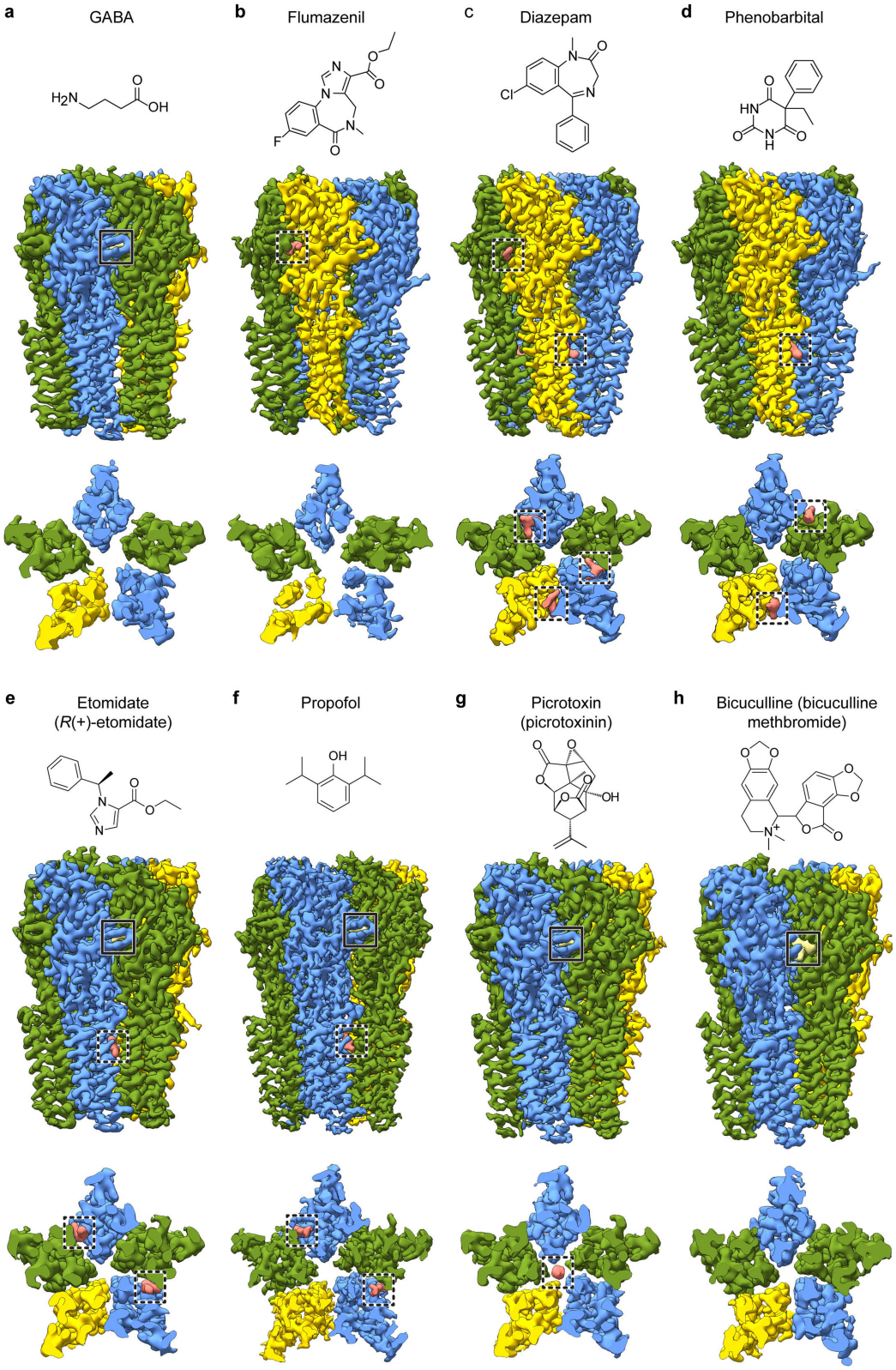
Extended Data Fig. 1 | Biochemistry, sample condition screening, and stability of atomistic molecular dynamics simulations in brain lipids. In 2018, our group reported the structure of the α1β2γ2 receptor in complex with GABA and flumazenil in detergent[21]. Although this initial study revealed details of the classical neurotransmitter and benzodiazepine binding sites, the structures showed an unanticipated asymmetric occluded state in the transmembrane region, where we observed the γ2 TMD collapsed into the pore or structurally disordered. Structures in complex with GABA[85] or with a nanobody modulator[86], also in detergent, exhibited very low resolution in the membrane domain that precluded detailed analysis. Structures of the α1β3γ2 receptor in lipid nanodiscs were reported more recently, with a well-ordered and approximately symmetric transmembrane domain[18,19]. We first sought to improve order and prevent collapse of the symmetric transmembrane domain (TMD) quaternary structure by optimizing lipid reconstitution of the GABA plus flumazenil receptor complex as a benchmark. **a**, Analytical size-exclusion chromatography of the α1β2γ2 receptor at different stages of preparation of the GABA plus flumazenil complex, which we used to benchmark the reconstitution approach: receptor in detergent, increasing in size after exchange into nanodiscs, then a further increase in size after addition of Fab. Inset, SDS–PAGE shows relatively pure nanodisc–Fab–receptor complex, which was used for grid preparation. **b**–**e**, TMD *z*-slices of 3D reconstructions from preparations with GABA, flumazenil and various membrane mimetics. Inset numbers are resolution values from the reconstructions and white dashed lines highlight subunit boundaries. **b** is from the dataset published in 2018[21]; **c** is from the sample purified in DDM supplemented with brain lipids, more symmetric but very low resolution; **d** is from protein purified in DDM supplemented with soy polar lipid extract (Avanti) and cholesteryl hemisuccinate (CHS, Anatrace) and exchanged in MSP1E3 nanodiscs containing soy lipids, highly asymmetric; **e** is the condition used to obtain the GABA plus flumazenil complex in this study. We applied this purification and nanodisc reconstitution approach to all other complexes. **f**, Results from atomistic molecular dynamics simulations validating the stability of these complexes in a brain-lipid environment, as well as differential dynamics in the presence of different ligands. After embedding our models in mixed membranes with expected brain-lipid proportions[87] and equilibrating with coarse-grained simulations[72], cholesterol and PtdIns(4,5)P_2 were found to accumulate at the protein surface, particularly at subunit interfaces (Supplementary Videos 9 and 10, respectively). Such interactions could contribute to the symmetrizing effect of brain lipids relative to detergent or other lipid mixtures. Subsequent quadruplicate 500-ns all-atom molecular dynamics simulations of all 8 structures reported in this work were largely stable, converging to ≤3 Å r.m.s.d. for all protein Cα atoms. This panel shows deviations from starting conformations (r.m.s.d., Å) of protein Cα atoms in α1β2γ2 receptor structures. Each trace represents one of four 500-ns replicates. **g**, An alternative conformation observed in multiple exploratory simulations of the flumazenil-bound structure (grey) with flumazenil removed. Within 200 ns, the γ M2-helix spontaneously translocates to block the pore (snapshot at 500 ns, coloured), supporting a flexible conformational repertoire for this subunit. Transition is tracked over time (red–blue) by the position of P-2′ in α and γ. **h**, Simulation results for propofol stability at all five interfacial TMD sites, with probability distributions at left, and raw data (n = 500 samples from 4 simulations, see Methods) plus box plots indicating sample median, interquartile range (25th–75th percentiles), minimum–maximum range, and outliers at right. Propofol was inserted at the α–β, α–γ and γ–β sites by symmetry superposition of the resolved β–α propofol. In quadruplicate simulations of >400 ns each, the inserted propofol molecules were not stably bound, sampling a broad distribution up to 8 Å r.m.s.d. from initial poses. By contrast, propofol at the β–α interfaces remained within 4 Å r.m.s.d. of its initial poses. Thus, simulations support a preference for propofol binding at the β–α interface over other interfaces.



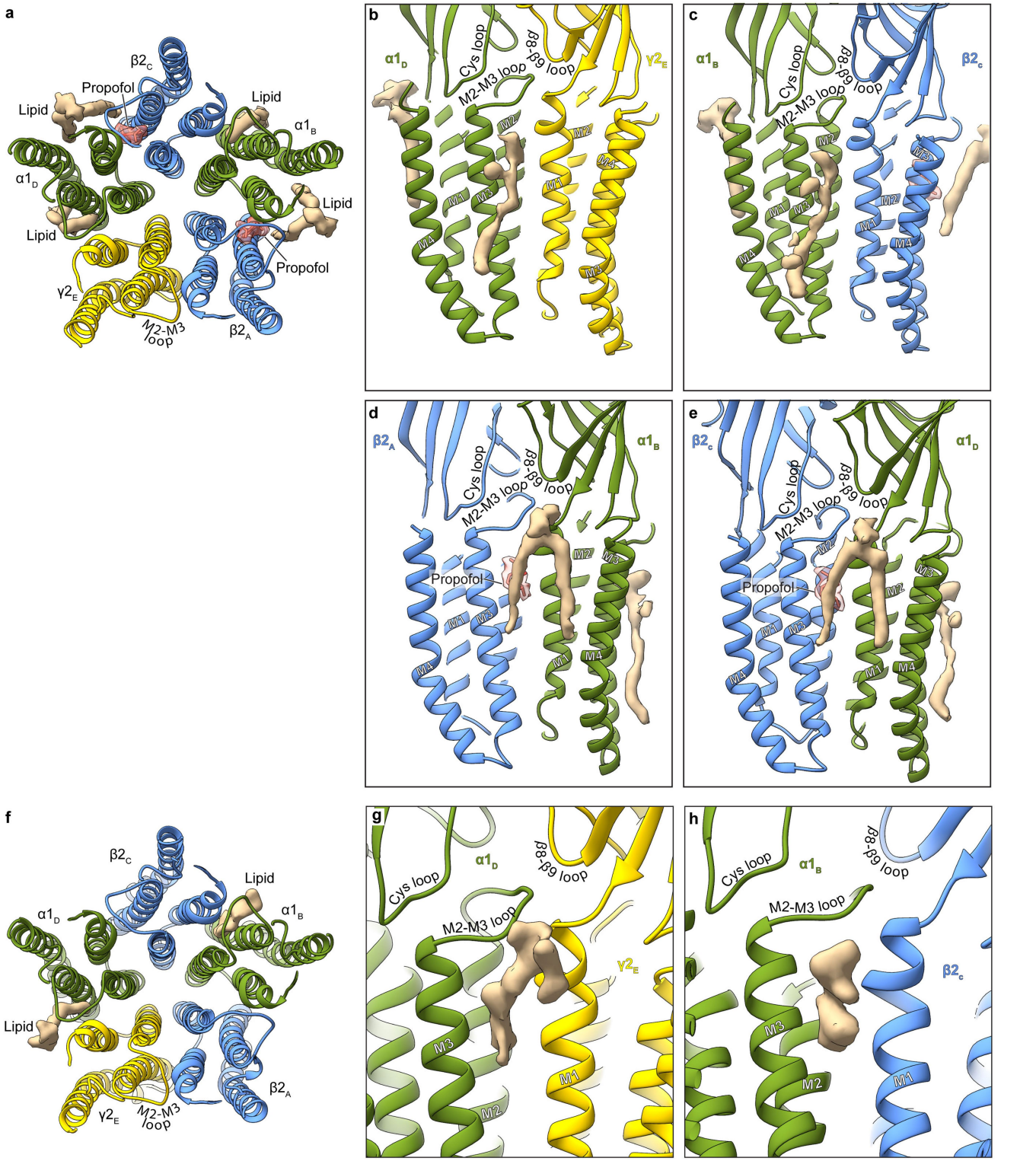
Extended Data Fig. 2 | Detailed cryo-EM processing flowchart for GABA plus flumazenil complex. a, A representative cryo-EM image. **b**, Projection images from the final selected 2D classes. **c**, 3D classification results; good classes selected for further processing are boxed in red and in lower row have TMD *z*-slices shown. Note fuzzy nanodisc appearance adjacent to γ2 subunit, consistent with conformational heterogeneity in this region. **d**–**f**, 3D maps from a second round of 3D classification (**d**), from which particle from four classes (red boxes) were selected and used to generate map shown in **e**. Signal subtraction and γ2 subunit focused 3D classification resulted in the map in **f**.



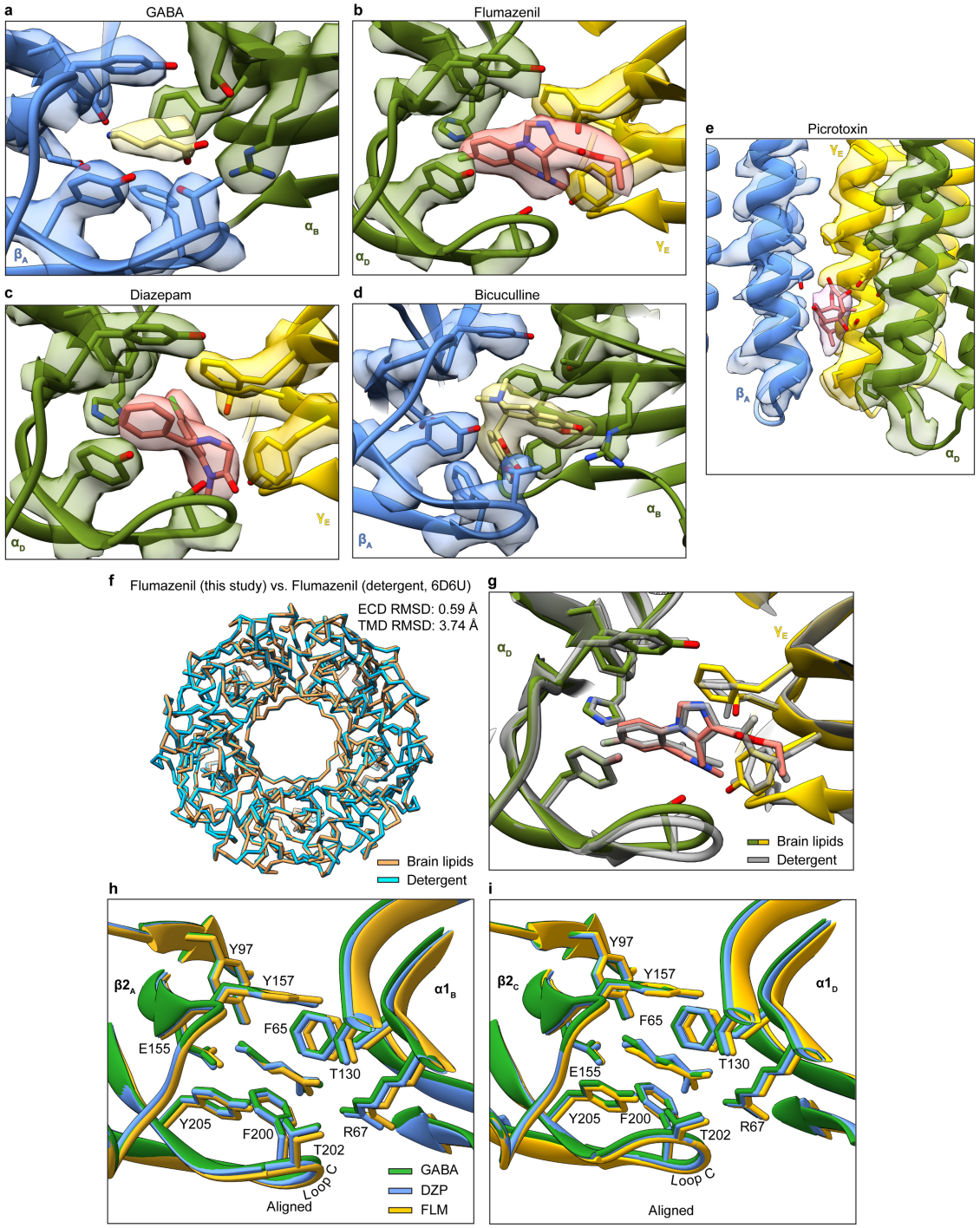
Extended Data Fig. 3 | Overall and local map resolution and global map–model agreement. For each structure, the sharpened map is coloured by local resolution, and map FSC (upper right) and map-model FSC (lower right) plots are shown. For the flumazenil complex, two maps were used in building: a higher resolution map that had weak γ-TMD density, and a lower resolution map with strong γ-TMD signal. Shown here, for this structure, is the lower resolution map with strong signal for the whole receptor. Both maps will be deposited for this flumazenil complex, and relevant statistics for these maps are shown in Extended Data Tables 1, 2.



Extended Data Fig. 4 | Map quality and ligand binding sites. a–h, Each panel shows a side view and a TMD slice from the experimental density map, accompanied by the chemical structure of the ligand in that complex. Note, GABA is present in all structures except the bicuculline complex. Solid boxes highlight GABA binding sites; dashed boxes highlight allosteric ligands (including picrotoxin) binding sites. Propofol binding sites at subunit interfaces in **f** are distinct from the intrasubunit sites identified initially in the prokaryotic GLIC channel[88], and similar in location but distinct in pose compared to the intersubunit site mutants of GLIC[89].



Extended Data Fig. 5 | Lipid interactions in TMD. a, An atomic model overview of the TMD sites for possible lipid binding in the GABA plus propofol complex; densities for putative lipids are shown in tan. A subset of these are consistent with those modelled as POPC in the α1β3γ2 structures[18,19]. **b**–**e**, Side views of lipid density at the different subunit interfaces. The lipid density maps shown were generated using the unsharpened map. **f**–**h**, Structure of the GABA plus diazepam complex. **f**, An atomic model overview of the TMD sites for possible lipid binding; densities for putative lipids are shown in tan. **g**, **h**, Side views of potential lipid density at the subunit interfaces.



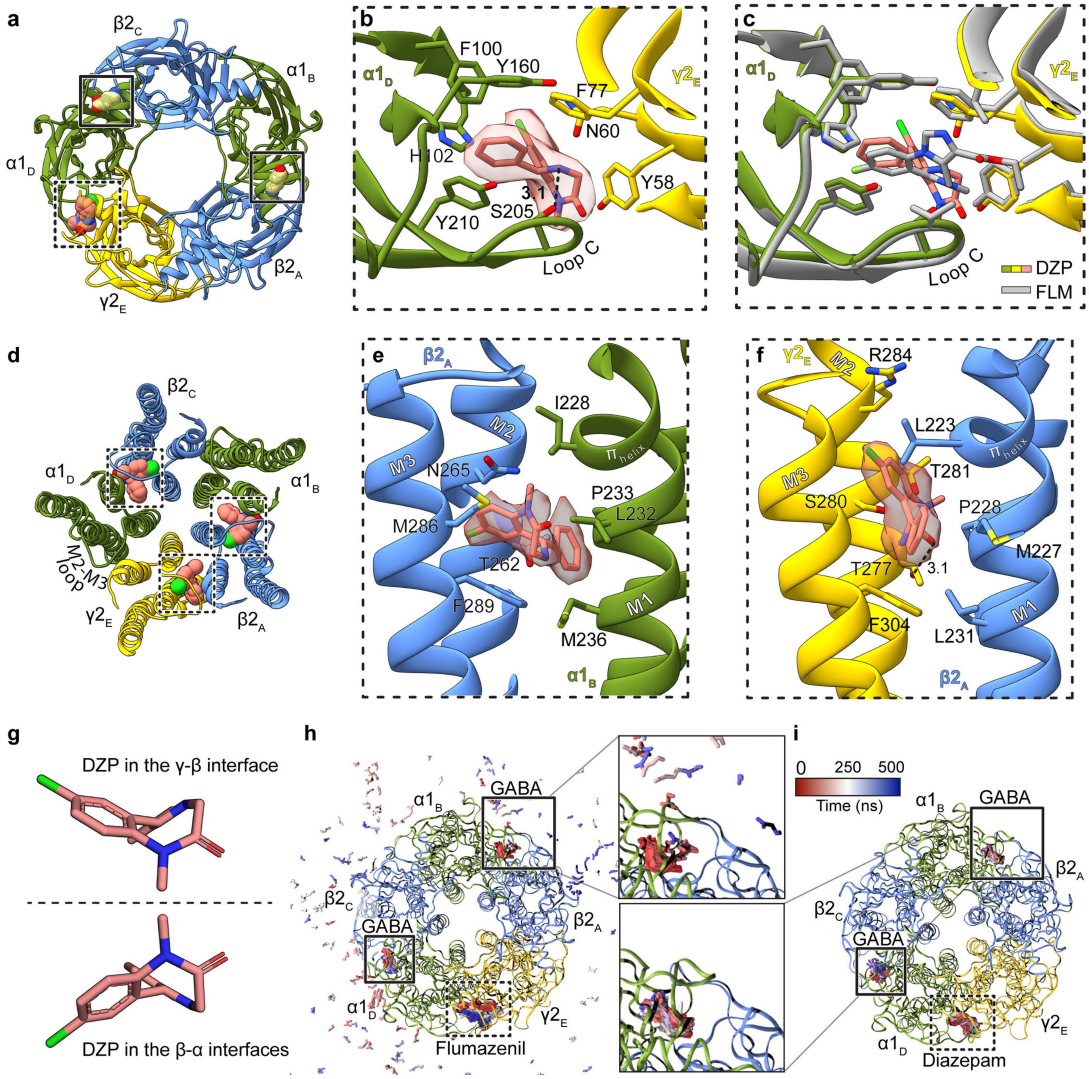
j

Complexes	Flumazenil			GABA			Diazepam		
Interfaces	All	ECD	TMD	All	ECD	TMD	All	ECD	TMD
α/γ (D/E)	2321.7/-28.1*	1217.7/-6.9	1011.6/-19.8	2511.7/-30.7	1351.9/-6.5	1073.8/-22.9	2461.8/-30.6	1289.0/-6.8	1086.0/-22.1
γ/β (E/A)	2250.0/-25.5	1492.2/-11.7	634.7/-13.0	2701.5/-28.1	1684.4/-9.7	919.0/-17.4	2769.1/-35.8	1572.7/-9.9	1102.6/-24.9
β/α (A/B)	2669.4/-33.4	1458.0/-14.5	1089.0/-17.7	2829.2/-42.6	1525.1/-17.1	1202.6/-23.6	2593.7/-37.3	1516.2/-17.4	1004.3/-18.9
α/β (B/C)	2367.5/-31.6	1480.4/-11.5	823.5/-19.8	2696.1/-33.3	1589.8/-12.3	1020.1/-20.0	2792.2/-36.4	1520.2/-12.6	1171.4/-23.4
β/α (C/D)	2632.7/-37.0	1513.0/-14.2	1021.7/-21.4	2740.0/-40.2	1532.9/-16.5	1113.4/-22.4	2642.2/-35.5	1510.7/-18.0	1069.1/-16.5

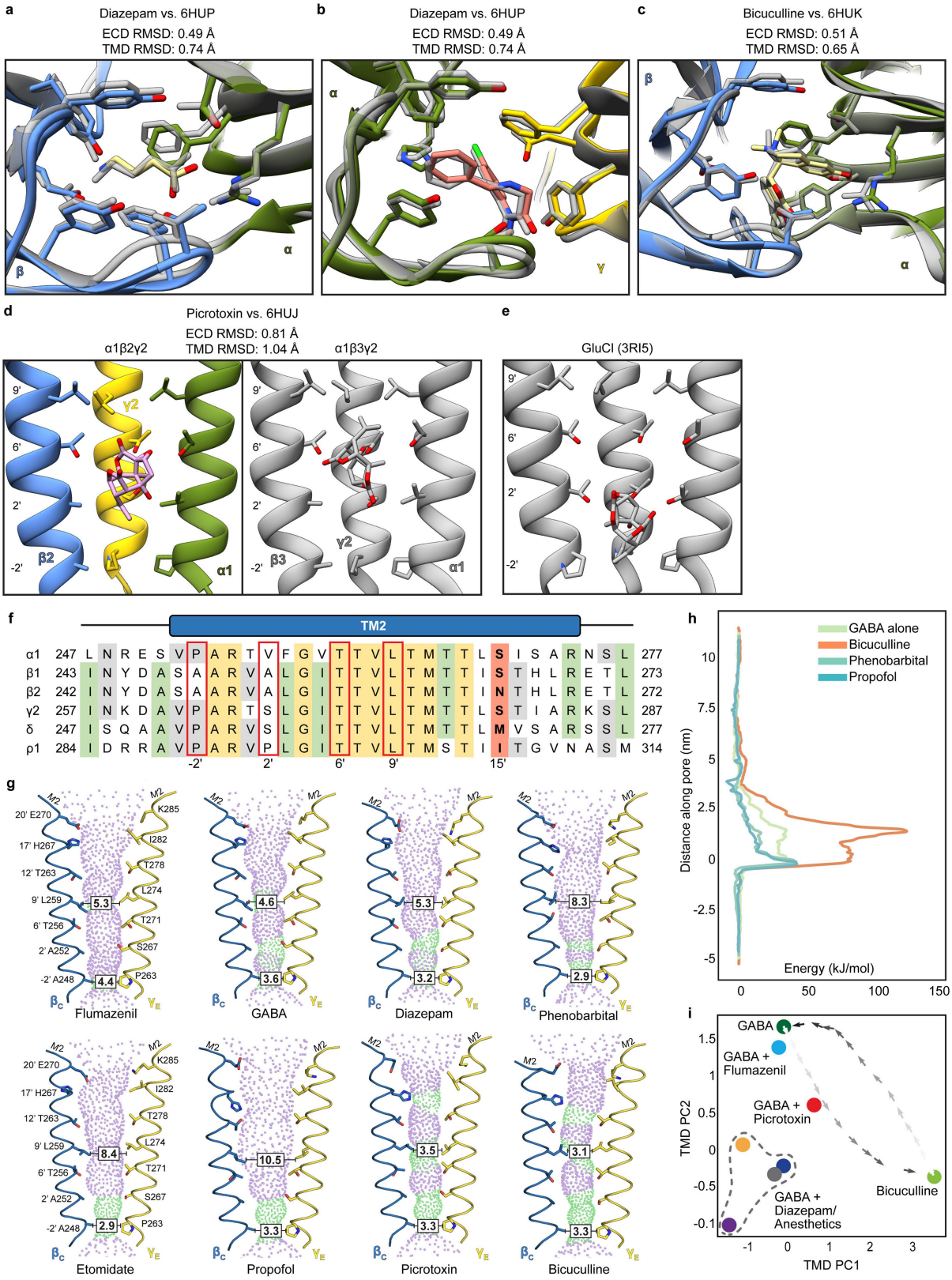
* Interface Area (Å²) / ΔiG (kcal/mol)

Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Representative map quality and model fit and structural analysis of GABA alone, diazepam and flumazenil complexes. Semitransparent surface is shown for central ligand and contacting side chains for **a**–**d**. **a**, The GABA site at chain A–B β–α interface in the GABA-alone structure. The two β–α GABA sites from the structure superimpose nearly perfectly and do not shed light on the differences in functional contributions found in electrophysiology studies with concatamers[48]. Structures of apo receptor may be essential in identifying structural differences in the two GABA sites. **b**, Flumazenil site at the α–γ interface. **c**, Diazepam at the same ECD interface. **d**, Bicuculline site at the same interface as in **a**. **e**, The picrotoxin site in TMD; here, density is shown for ligand and all nearby protein structure elements. **f**, Superposition of two GABA plus flumazenil complexes, one from the detergent condition[21] and one from this study in brain lipids, to illustrate absence of differences in backbone conformation. Note, loops that interact with the TMD do vary in conformation. **g**, Detail of flumazenil site from the superposition in **f**. **h**, **i**, Superpositions of three structures from the current study: GABA alone, GABA plus diazepam and GABA plus flumazenil, focused on the two GABA-binding sites. **j**, Calculated interface areas and interaction energies for each subunit pair, for each of the benzodiazepine-related structures.

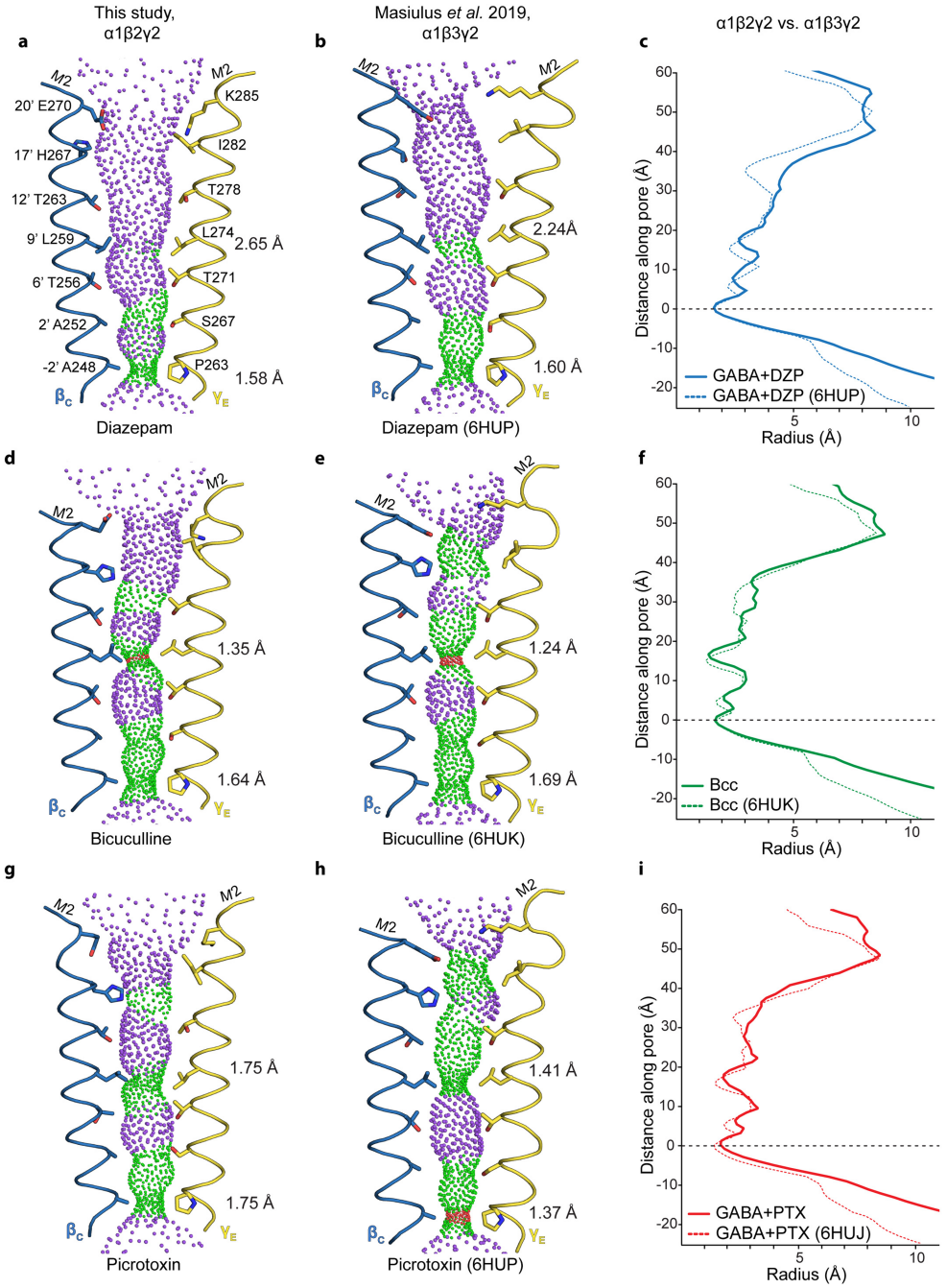


Extended Data Fig. 7 | Agonist and benzodiazepine complexes. a–**c**, ECD binding sites viewed from the synaptic perspective, **a**, Overview of the diazepam complex. **b**, Position of diazepam with ligand map quality shown; side chains shown for residues contacting diazepam. **c**, Superposition of flumazenil and diazepam complexes. **d**, The three TMD sites identified for diazepam. **e**, **f**, Binding site details for diazepam at the β–α and γ–β interfaces. **g**, The two enantiomeric conformations of diazepam identified in the TMD sites. **h**, **i**, Snapshots from molecular dynamics simulations viewed from the extracellular side. Extracellular GABA and benzodiazepines are shown as sticks, coloured by frame (red–blue scale). **h**, Flumazenil-bound simulation with GABA in the upper site unbinding within 100 ns (pink–blue peripheral sticks). **i**, Diazepam-bound simulation with GABA retained in both orthosteric sites. Subunit subscripts denote chain ID. Stick representation is shown for residues within the van der Waals contact range.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Ligand site comparisons among α1β2γ2, α1β3γ2 and GluCl structures, and panel of pore conformations. a, **b**, Superpositions of the GABA and diazepam ECD binding sites from the α1β2γ2 receptor (this study; subunits and ligands are coloured) and the α1β3γ2 receptor (in grey)[19], respectively. **c**, A superposition similar to those in **a** and **c** but for the bicuculline complexes (*N*,*N*-dimethyl is the higher-affinity form from this study; bicuculline (single *N*-methyl) for α1β3γ2 in grey). **d**, **e**, Comparison of picrotoxin binding sites from three structures: this study, the α1β3γ2 structure and GluCl[90]. The results suggest that picrotoxin can bind to multiple conformations at different depths of the pore. GluCl is most widely open and picrotoxin binds most deeply; in that study, picrotoxin was used as a probe for an open-state conformation[90]. The pore is more tightly closed in α1β3γ2 than in α1β2γ2, which may allow picrotoxin to bind more deeply in the latter structure. In GluCl and in α1β2γ2, the picrotoxin isoprenyl tail orients towards the cytosol; in α1β3γ2, tail orients towards extracellular surface. This orientation allows in GluCl for favourable interactions between the 'basket' oxygens and the polar 2′ residues. The α1β2/β3γ2 receptors are more hydrophobic at the 2′ position, which might also explain favourable positioning of picrotoxin higher in the pore, where in the α1β2γ2 structure these oxygens are likely to make hydrogen-bonding interactions with conserved 6′ threonine hydroxyls. **f**, A sequence alignment of $GABA_A$ subunit M2 helices. Red boxes highlight residues potentially important in picrotoxin binding; in bold are the 15′ residues that have a role in anaesthetic selectivity and sensitivity. **g**, Pore conformational states for all ligand complexes, with opposing β1 and γ2 M2 α-helices shown as ribbons with pore-lining side chains shown as sticks. Purple and green spheres illustrate shape of the pore. Boxed distances in the pore are diameters at the desensitization gate (−2′) and resting gate (9′) positions. **h**, Free energies for chloride ion permeation along the pore axis (cytoplasmic side down, with −2′ gate at 0 nm), for representative α1β2γ2 complexes. Overlaid plots show the energy barrier at the 9′ hydrophobic gate (around 2 nm) in the bicuculline complex (orange) to be partially relieved in the GABA complex (green), and further relieved in complexes with GABA + phenobarbital or GABA + propofol (light or dark blue, respectively). **i**, All α1β2γ2 structures reported in this work ($n = 8$ independent structures), plotted along dominant principal components calculated for the TMD. Snapshots of a simulated transition[78] between the GABA and bicuculline complexes (light-to-dark crosses) show that the GABA + picrotoxin complex maps along this pathway. GABA + diazepam and intravenous-anaesthetic-bound structures (GABA + diazepam, dark blue; etomidate, grey; phenobarbital, orange; propofol, purple) cluster at the lower left, distinct from GABA-alone or flumazenil- or inhibitor-bound states.



j

Average buried Interface area (Å²) at ligand-bound subunits

Complexes	Area(Å²)
GlyR α1, strychinine - 3JAD	2121.3
GlyR α3, strychinine - 5CFB	2231.0
GluCl, APO - 4TNV	2109.5
$GABA_AR$, bicuculline - 6HUK	3002.3
$GABA_AR$, bicuculline - this study	2679.4

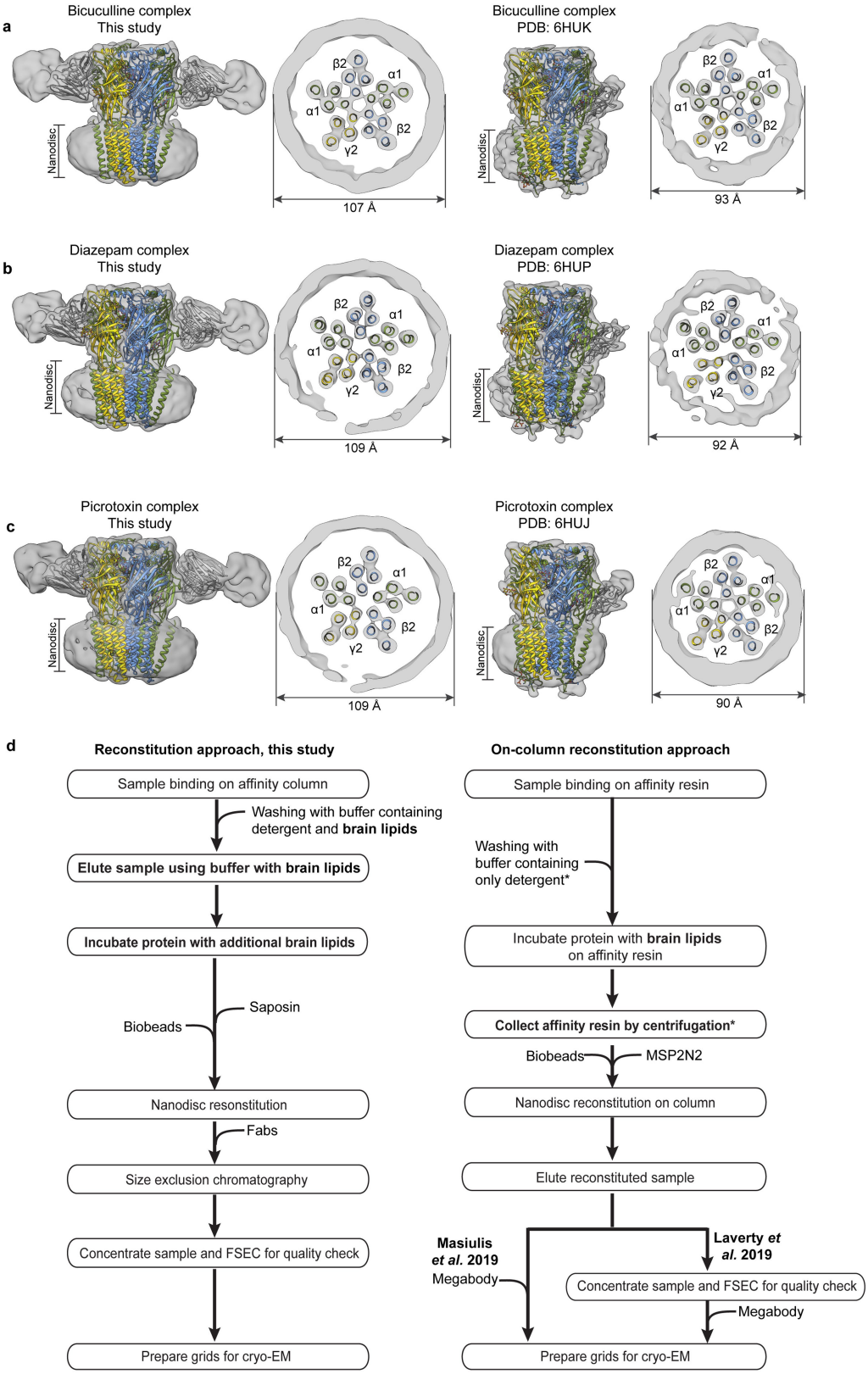
k

Interface area (Å²) in the TMD

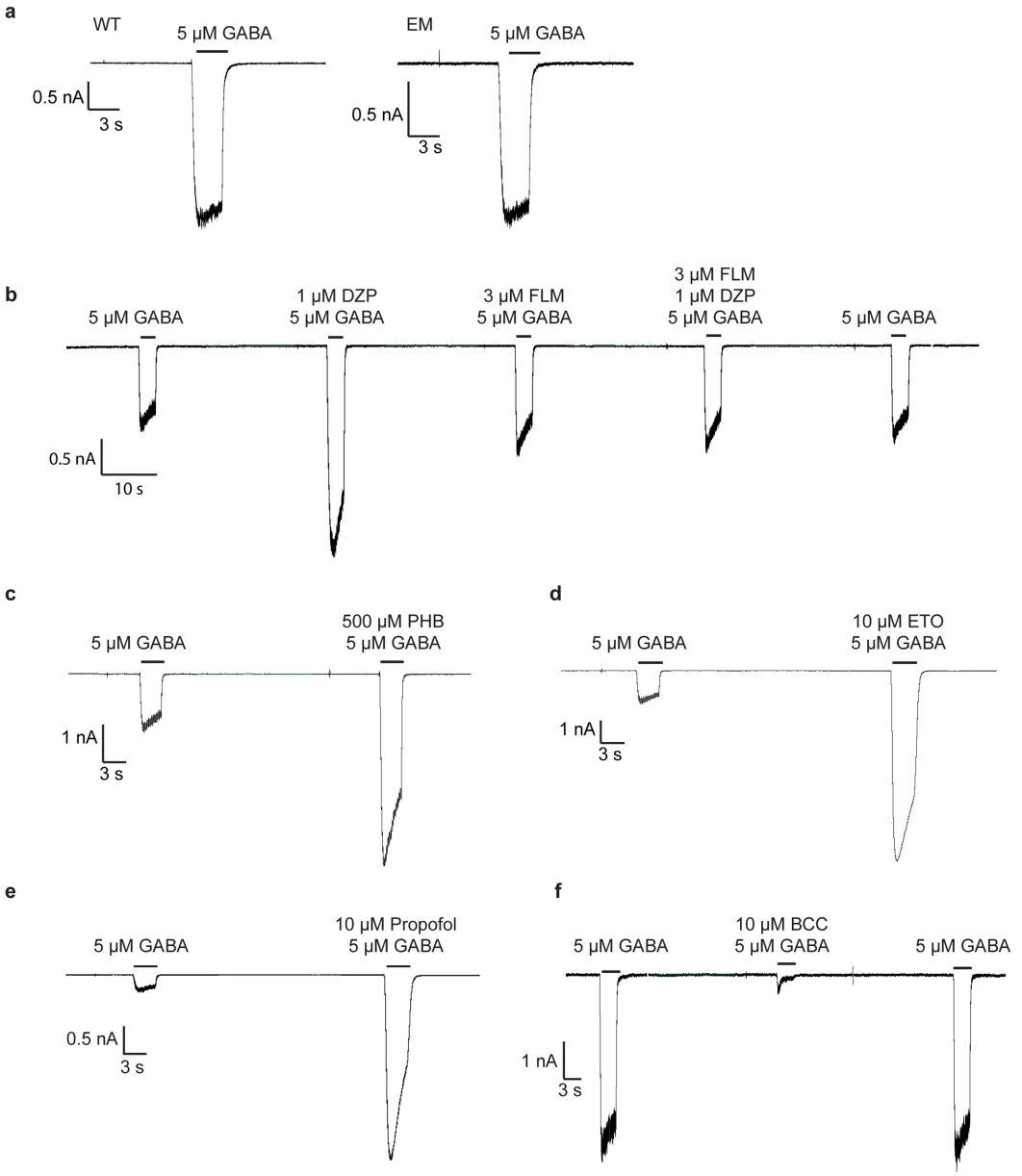
Complexes	Diazepam		Picrotoxin		Bicuculline	
Interfaces	In this study	6HUP	In this study	6HUJ	In this study	6HUK
α/γ	1086.0	1061.2	1063.2	1196.1	1086.6	1148.6
γ/β	1102.6	1163.3	1210.4	1178.5	1133.2	1209.3
β/α	1004.3	1368.7	1218.1	1592.6	1335.4	1602.0
α/β	1171.4	1099.4	1089.3	1203.4	1126.8	1175.7
β/α	1069.1	1269.5	1267.7	1508.6	1330.0	1555.1
Sum	5433.4	5962.1	5848.7	6679.2	6012.0	6690.7

Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Ion-pore conformation and TMD subunit interface packing in α1β2γ2 compared with α1β3γ2 structures. a, **b**, Pore conformations for α1β2γ2 (this study) (**a**) and α1β3γ2[19] (**b**) structures bound by GABA plus diazepam, with opposing β1 and γ2 M2 α-helices shown as ribbons and pore-lining side chains shown as sticks. Purple and green spheres illustrate the shape of the pore; purple is for radii >2.8 Å; green is 1.4–2.8 Å; red is < 1.4 Å. Distances on the right side of pore are radii at the desensitization gate (−2′) and resting gate (9′) positions. **c**, A comparison of these two structures in the form of a pore radius versus distance along the pore plot. Structures were aligned at $y = 0$ at the level of the −2′ desensitization gate. **d**–**f** and **g**–**i** make the same comparisons, but for the bicuculline (**d**–**f**) and GABA plus picrotoxin (**g**–**i**) complexes. **j**, Comparison of the interface area buried per subunit interface (Å^2, ECD+TMD) for representative anion-selective receptors; top three are homopentamers for which the area given is the average from all interfaces, whereas for the two bicuculline structures the area comes from the average of the two β−α interfaces. Comparison is limited to anion-selective receptors owing to the absence of ordered intracellular domains; eukaryotic cation-selective receptors contain intracellular domains that contribute to interface surface area. **k**, Buried TMD subunit interface areas between pairs of $GABA_A$ receptor structures, to illustrate tighter packing in the α1β3γ2 receptor structures.



Extended Data Fig. 10 | Nanodisc sizes correspond to the lipid ratio used in reconstitution. a–**c**, Comparison of experimental EM maps (with docked structures), low-pass-filtered to 10 Å resolution, between matched α1β2γ2 and α1β3γ2 ligand complexes. **d**, Comparison of the reconstitution approach from the current study with the on-column approach used to obtain the α1β3γ2 receptor structures[18,19]. Asterisks indicate steps we propose give rise to the observed different nanodisc sizes: washing with lipid-free detergent buffer removes lipids, and the step of collecting affinity resin by centrifugation removes excess lipids, such that when the MSP2N2 scaffold and Bio-Beads are added, there are no extra lipids to fill the large scaffold.



Extended Data Fig. 11 | Example electrophysiological recordings with cryo-EM construct. All recordings were made in whole-cell voltage-clamp mode at −75 mV with transiently-transfected HEK cells. **a**, Wild-type, full-length receptor compared to cryo-EM construct, response to application of GABA. All remaining recordings are with the EM construct. **b**, A representative response is shown for application of GABA, then GABA plus diazepam, then GABA plus flumazenil, then GABA plus diazepam plus flumazenil. **c**, Application of GABA, then GABA plus phenobarbital. **d**, Application of GABA, then GABA plus etomidate. **e**, Application of GABA, then GABA plus propofol. **f**, Application of GABA, then GABA plus the methylated form of bicuculline. The patch-clamp experiments were repeated 3 times independently.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics for bicuculline methbromide, GABA + propofol, GABA + etomidate and GABA + phenobarbital complexes

	Bicuculline methbromide (EMDB-22031) (PDB 6X3S)	GABA+ Propofol (EMDB-22032) (PDB 6X3T)	GABA+ Etomidate (EMDB-22034) (PDB 6X3V)	GABA+ Phenobarbital (EMDB-22035) (PDB 6X3W)
Data collection and processing				
EM Facility	UTSW	PNCC	UTSW	HMS
Magnification	105 K	22.5 K	105 K	105 K
Voltage (kV)	300	300	300	300
Electron exposure (e–/Å^2)	85.05	50.00	66.07	69.59
Defocus range (μm)	-1.8 to -2.8	-1.8 to -2.8	-1.8 to -2.8	-1.8 to -2.8
Pixel size (Å)	0.833	1.035	0.833	0.825
Symmetry imposed	C1	C1	C1	C1
Initial particle images (no.)	1,219,070	3,052,289	1,972,936	1,076,196
Particle images after 2D classification	815,729	926,936	719,534	513,431
Final particle images (no.) (%)‡	80,103 (9.8)	158,159 (17.1)	124,310 (17.3)	145,958 (28.4)
Map resolution (Å)	3.12	2.55	3.47	3.14
FSC threshold	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.7 – 4.2	2 – 2.8	3 – 4.0	2.4 – 3.8
Refinement				
Initial model used (PDB code)	6X3Z	6X3Z	6X3Z	6X3Z
Model resolution (Å)	3.31	2.60	3.56	3.24
FSC threshold	0.5	0.5	0.5	0.5
Model resolution range (Å)	n.a.	n.a.	n.a.	n.a.
Map sharpening *B* factor (Å^2)	-86	-61	-100	-102
Model composition				
Non-hydrogen atoms	17,407	17,416	17,426	17,399
Protein residues	2,121	2,121	2,121	2,121
Ligands	23	27	27	25
B factors (Å^2)				
Protein	43.00	41.53	29.54	35.92
Ligand	52.96	42.31	40.90	45.94
R.m.s. deviations				
Bond lengths (Å)	0.004	0.004	0.005	0.008
Bond angles (°)	0.599	0.586	0.568	1.133
Validation				
MolProbity score	1.79 (100th %)	1.77 (99th %)	1.72 (100th %)	1.59 (100th %)
Clashscore	9.42 (96th %)	5.88 (99th %)	7.33 (100th %)	5.41 (100th %)
Poor rotamers (%)	0.05	2.28	0.58	0.85
Ramachandran plot				
Favored (%)	95.72	96.96	95.48	95.67
Allowed (%)	4.28	3.04	4.52	4.33
Disallowed (%)	0.00	0.00	0.00	0.00

‡Percentage of particles in final reconstitution compared to after 2D classification.

Extended Data Table 2 | Cryo-EM data collection, refinement and validation statistics for GABA, GABA + diazepam, GABA + flumazenil and GABA + picrotoxin complexes

	GABA (EMDB-22037) (PDB 6X3Z)	GABA+ Diazepam (EMDB-22036) (PDB 6X3X)	GABA+ Flumazenil (EMDB-22033) (PDB 6X3U)	GABA+ Picrotoxin (EMDB-22038) (PDB 6X40)
Data collection and processing				
EM Facility	HMS	UTSW	UTSW	HMS
Magnification	105 K	105 K	165 K	105 K
Voltage (kV)	300	300	300	300
Electron exposure (e–/Å^2)	63.79	63.04	50.28	62.91
Defocus range (μm)	-1.8 to -2.8	-1.8 to -2.8	-0.6 to -2.1	-1.4 to -2.3
Pixel size (Å)	0.825	0.833	0.84	0.825
Symmetry imposed	C1	C1	C1	C1
Initial particle images (no.)	1,705,334	2,868,814	1,072,111	1,847,538
Particle images after 2D classification	871,313	826,254	810,710	1,550,272
Final particle images (no.) (%)‡	171,838 (19.7)	297,028 (35.9)	260,276 / 62,364 (32.1 / 7.7)	165,494 (10.7)
Map resolution (Å)	3.23	2.92	3.20 / 3.49	2.86
FSC threshold	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.5 – 3.9	2 – 3.5	3 – 4.5	2 – 3.2
Refinement				
Initial model used (PDB code)	6U6D + 4COF	6X3Z	6X3Z	6X3Z
Model resolution (Å)	3.34	3.02	3.70	2.99
FSC threshold	0.5	0.5	0.5	0.5
Model resolution range (Å)	n.a.	n.a.	n.a.	n.a.
Map sharpening *B* factor (Å^2)	-109	-113	-125 / -120	-94
Model composition				
Non-hydrogen atoms	17,365	17,470	17,387	17,411
Protein residues	2,121	2,121	2,121	2,121
Ligands	23	29	24	26
B factors (Å^2)				
Protein	62.75	40.07	24.15	62.21
Ligand	66.81	46.42	31.71	60.63
R.m.s. deviations				
Bond lengths (Å)	0.007	0.006	0.005	0.007
Bond angles (°)	1.069	1.024	0.998	1.092
Validation				
MolProbity score	1.59 (100th %)	1.49 (100th %)	1.47 (100th %)	1.68 (100th %)
Clashscore	5.44 (100th %)	4.14 (100th %)	4.05 (100th %)	7.16 (98th %)
Poor rotamers (%)	0.69	0.53	0.32	0.42
Ramachandran plot				
Favored (%)	95.77	95.82	95.91	95.82
Allowed (%)	4.23	4.18	4.09	4.18
Disallowed (%)	0.00	0.00	0.00	0.00

‡Percentage of particles in final reconstitution compared to after 2D classification.

nature research

Corresponding author(s): Ryan E. Hibbs

Last updated by author(s): May 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	pClamp10, SerialEM
Data analysis	GraphPad Prism 8, SerialEM, Coot 0.8.9.2-pre-revision-7884, Phenix 1.17.1-3660, Relion 3.0.8_cu8.0, Relion 3.1-beta_cu8.0, MotionCor2, GCTF, CTFFIND4, crYOLO, Chimera 1.14, ResMap, Pymol 2.3.2_79, Hole2, GROMACS 2019.4, CHARMM-GUI, VMD, MDAnalysis, MDTraj, eBDIMS

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability: Atomic model coordinates for bicuculline, GABA + propofol, GABA + flumazenil, GABA + etomidate, GABA + phenobarbital, GABA + diazepam, GABA, and GABA + picrotoxin–bound structures have been deposited in the Protein Data Bank with accession codes 6X3S, 6X3T, 6X3U, 6X3V, 6X3W, 6X3X, 6X3Z and 6X40, respectively, and the cryo-EM density maps have been deposited in the Electron Microscopy Data Bank with accession codes EMD-22031, EMD-22032, EMD-22033, EMD-22034, EMD-22035, EMD-22036, EMD-22037 and EMD-22038, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to estimate appropriate sample size.
Data exclusions	RELION-based 2D and 3D classification of single particle datasets were used to remove false positives and particles with disordered domains.
Replication	Attempts to replicate/reproduce the data were successful as detailed in the MD simulation and electrophysiology replicates.
Randomization	EM particle sets were randomly split for the purposes of estimating overall resolution.
Blinding	No blinding was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	1F4 Fab fragment was used in EM studies, published previously in PMID: 29950725.
Validation	Binding was validated by gel filtration and structural analysis; sequence was verified by sequencing of hybridoma line and by experimental density map.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC CRL-3022; ATCC CRL-1711
Authentication	Purchased from commercial supplier. Not further tested.
Mycoplasma contamination	Not tested.
Commonly misidentified lines (See ICLAC register)	n/a

Author Correction: Molecular heterogeneity drives reconfigurable nematic liquid crystal drops

https://doi.org/10.1038/s41586-020-2659-0

Correction to: *Nature* https://doi.org/10.1038/s41586-019-1809-8

Published online 18 December 2019

Check for updates

Wei-Shao Wei, Yu Xia, Sophie Ettinger, Shu Yang & A. G. Yodh

In this Article, in the 'Calculation of system free energy' section of the Methods, there were omissions in equation (7). The equation should read:

"Or, if $K_{11} < K_{33}$,

$$F_f = \pi K_{11}\left[2 + \frac{k}{\sqrt{k-1}}\tan^{-1}\sqrt{k-1} - \frac{k}{\sqrt{k-1}}\tan^{-1}\left(\frac{\sqrt{k-1}}{\sigma}\right) - \frac{K_{24}}{K_{11}}\right]L + \gamma 2\pi rL$$

Or, if $K_{11} > K_{33}$,

$$F_f = \pi K_{11}\left[2 + \frac{k}{\sqrt{1-k}}\tanh^{-1}\sqrt{1-k} - \frac{k}{\sqrt{1-k}}\tanh^{-1}\left(\frac{\sqrt{1-k}}{\sigma}\right) - \frac{K_{24}}{K_{11}}\right]L + \gamma 2\pi rL \quad (7)''$$

instead of:

"Or, if $K_{11} < K_{33}$,

$$F_f = \pi K_{11}\left[2 + \frac{k}{\sqrt{k-1}}\tan^{-1}\sqrt{k-1} - \frac{k}{\sqrt{k-1}}\tan^{-1}\left(\frac{\sqrt{k-1}}{\sigma}\right) - \frac{K_{24}}{K_{11}}\right]$$

Or, if $K_{11} > K_{33}$:

$$F_f = \pi K_{11}\left[2 + \frac{k}{\sqrt{1-k}}\tanh^{-1}\sqrt{1-k} - \frac{k}{\sqrt{1-k}}\tanh^{-1}\left(\frac{\sqrt{1-k}}{\sigma}\right) - \frac{K_{24}}{K_{11}}\right] \quad (7)''$$

All calculations and conclusions presented in the Article were carried out using the correct equations and are unaffected by these changes. The original Article has been corrected online.

Publisher Correction: A dynamically cold disk galaxy in the early Universe

https://doi.org/10.1038/s41586-020-2693-y

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2572-6

Published online 12 August 2020


Check for updates

F. Rizzo, S. Vegetti, D. Powell, F. Fraternali, J. P. McKean, H. R. Stacey & S. D. M. White

In this Article, owing to a typesetting error, the published online date of 8 December 2020 is incorrect in the print version; it appears correctly online as 12 August 2020.

Publisher Correction: The effect of large-scale anti-contagion policies on the COVID-19 pandemic

https://doi.org/10.1038/s41586-020-2691-0

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2404-8

Published online 8 June 2020


Check for updates

Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, Peiley Lau, Jaecheol Lee, Esther Rolf, Jeanette Tseng & Tiffany Wu

In this Article, owing to a typesetting error, the published online date of 6 August 2020 is incorrect in the print version; it appears correctly online as 8 June 2020.

Publisher Correction: Temperature-dependent growth contributes to long-term cold sensing

https://doi.org/10.1038/s41586-020-2694-x

Correction to: *Nature* https://doi.org/10.1038/s41586-020-2485-4

Published online 15 July 2020


Check for updates

Yusheng Zhao, Rea L. Antoniou-Kourounioti, Grant Calder, Caroline Dean & Martin Howard

In the HTML version of this Article, the present address 'Department of Biology, University of York, York, UK' was erroneously associated with author Caroline Dean. This author's affiliation should be 'John Innes Centre, Norwich Research Park, Norwich, UK' only. The Article has been corrected online.

Work

Your story

Send your careers story to: naturecareerseditor@nature.com

GETTY

PANDEMIC DARKENS POSTDOCS' WORK AND CAREER HOPES

Nature's survey of this key segment of the scientific workforce paints a gloomy picture of interrupted research and anxiety about the future.

By Chris Woolston

Eight out of ten postdoctoral researchers say that the global coronavirus pandemic has hampered their ability to conduct experiments or collect data. More than half are finding it harder to discuss their research ideas or share their work with their laboratory head or colleagues, and nearly two-thirds believe that the pandemic has negatively affected their career prospects, according to *Nature*'s first-ever survey of postdocs worldwide (see 'Disruption and distress').

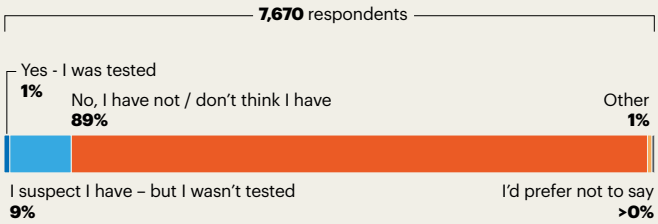
The pandemic has shuttered or reduced the output of academic labs globally, slashed institutional budgets and threatened the availability of grants, fellowships and other postdoctoral funding sources. The fallout adds up to a major challenge for a group of junior researchers who were already grappling with limited funds, intense job competition and career uncertainties (see Editorial, page 160).

Nature's self-selected survey, which ran from mid-June to the end of July and drew responses from 7,670 postdocs working in academia, included detailed questions on the impact of COVID-19 on the global postdoctoral community. Follow-up interviews with selected respondents and hundreds of free-text comments (see 'The situation is grim' for a selection) filled in an unsettled, precarious picture

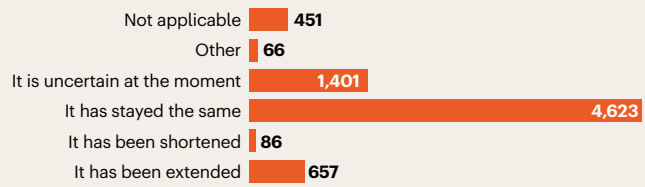
DISRUPTION AND DISTRESS

Nature's first-ever survey of more than 7,600 postdoctoral researchers worldwide uncovered great apprehension and uncertainty around the coronavirus pandemic's effect on respondents' current posts and career aspirations, and gaps in their ability to conduct research, maintain and secure funding, and communicate with their laboratory head and colleagues.

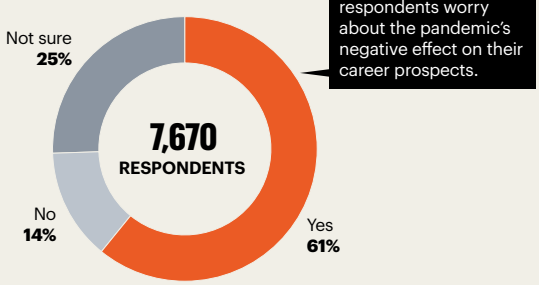
Q: Have you had COVID-19?



Q: Has your fellowship or term been extended because of COVID-19?

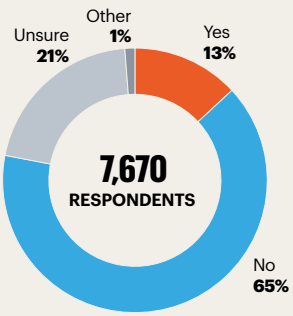


Q: Do you think the coronavirus pandemic has negatively affected your career prospects?



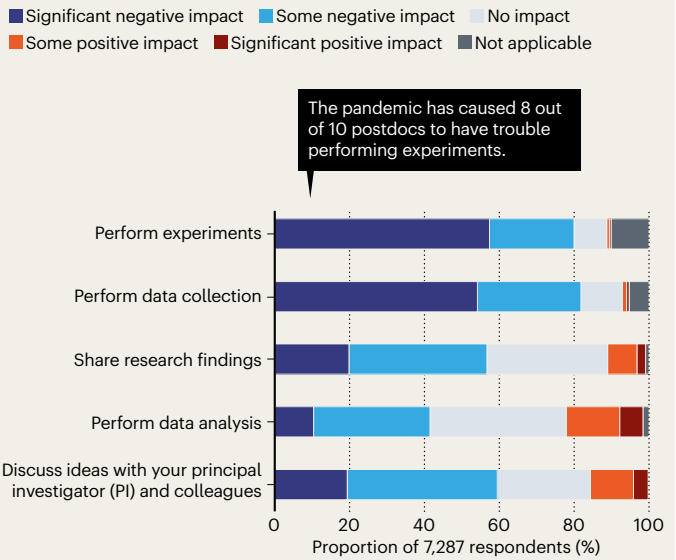
Almost two-thirds of respondents worry about the pandemic's negative effect on their career prospects.

Q: Do you believe you've lost a postdoc or post-postdoc job offer because of COVID-19?



More than one-fifth of postdocs aren't sure if the pandemic caused them to lose a job offer, whereas another 13% are certain it was the reason for their lost offer.

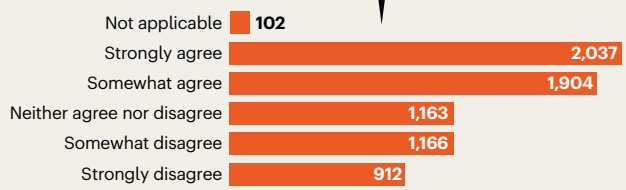
Q: How has COVID-19 affected your ability to do the following?



The pandemic has caused 8 out of 10 postdocs to have trouble performing experiments.

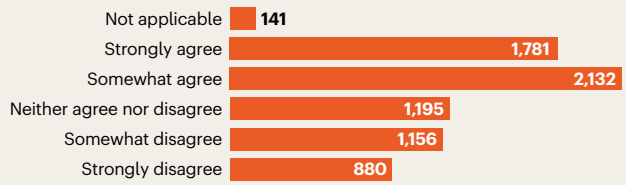
Nearly 6 out of 10 postdocs have trouble discussing ideas with their lab head and colleagues because of the pandemic.

Q: My supervisor/PI has done all I believe they can or should do to support me at this time.



More than half of postdocs feel supported by their principal investigator during the pandemic.

Q: My supervisor/PI has provided clear guidelines on how they will support me to manage any changes in my ability to work.



More than one-quarter of postdocs report a lack of guidance from their principal investigator around their inability to work because of the pandemic.

of postdoctoral research in the era of coronavirus. "The [pandemic] has compounded the pressures that postdocs were already under," says Hannah Wardill, a cancer researcher at the South Australian Health and Medical Research Institute in Adelaide, in an interview.

The survey, created together with Shift Learning, a market-research company based in London, was advertised on nature.com, in Springer Nature digital products and through e-mail campaigns. It was offered in English, Mandarin Chinese, Spanish, French and Portuguese. The data set relating to the COVID-19 responses is available at go.nature.com/34wrre1. The full results are currently being analysed and will be released in November.

Uncertain job prospects

One per cent of respondents say that they have been diagnosed with COVID-19, and another 9% suspect that they have had the infection but were never tested. But concerns go far beyond the presence or absence of the virus. Some 61% of respondents say that the pandemic has negatively affected their career prospects, and another 25% say that its cumulative effects on their career remain uncertain.

Worries about one's professional future are especially widespread in South America, where 70% of respondents say their careers have already suffered since the start of the pandemic. A biochemist in Brazil used the survey's comment section to share her own concerns. She notes that postdoctoral contracts in her country usually last for just one or two years, and extensions are far from guaranteed,

ANDREAS FECHNER

Cancer researcher Hannah Wardill had to cut short a promising project abroad to return to her position in Australia.

creating a tenuous situation for researchers who were probably already struggling to get by. "Here, we live in a reality where PhDs need to sell food on the street to support themselves financially, as most are unable to obtain scholarships or jobs," she wrote.

Julieth Caro, a physicist at the Federal University of Rio de Janeiro in Brazil, worries that the Brazilian government might shorten the length of her scholarship in a cost-cutting move. "The pandemic just makes me remember that science is not important to the government," she says. She adds that her scholarship prohibits her from taking a job outside her field. With few physics jobs available, she teaches experimental physics as an unpaid volunteer.

Belief that the pandemic had already negatively affected career prospects were also common in North and Central America (68%), Australasia (68%), Asia (61%), Africa (59%) and Europe (54%). In China, where the virus was first detected, 54% of respondents said their career had already suffered and 25% said they weren't sure.

Perceived impacts varied by area of study. Slightly less than half of researchers in computer science and mathematics thought that their career prospects had suffered, compared with 68% of researchers in chemistry, 67% in ecology and evolution, and 60% in biomedicine.

The impact of the pandemic has now joined the list of the top concerns in the minds of postdocs. Asked to name the three primary challenges to their career progression, 40% of respondents point to the economic impact of COVID-19, nearly two-thirds (64%) note the competition for funding, and 45% point to the lack of jobs in their field.

For those hoping to secure faculty jobs in 2020, the pandemic — and the widespread hiring freezes that have followed — could hardly have come at a worse time. A bioengineer in Germany used the comment section to explain his situation. "I had verbal faculty offers from multiple universities. During the COVID-19 pandemic, they practically froze the hiring but they did not even update me about it."

"Some 61% of respondents said that the pandemic had negatively affected their career prospects."

An HIV researcher in the United States who started looking for tenure-track positions this year comments that the pandemic may be a breaking point. "It's impossible to understate the impact that COVID-19 will have on our careers," he writes. "I'd like to stay in academia, but that may no longer be possible."

Thirteen per cent of respondents say they have already lost a postdoc job or an offer of one as a result of the pandemic, and 21% suspected the virus had wiped out a job but weren't sure. More than one-third of researchers in South America report already losing a job, compared with 11% in Europe and 12% in North and Central America.

Sixty per cent of respondents are currently working abroad, a circumstance that only amplifies the pandemic's potential impact. On top of everything else, many worry about the pandemic's effect on their visas and their ability to stay in their new country. A biochemist from India who is currently working in the United States wrote, "I'm on a visa that will expire in January 2021. Because of the COVID lockdown, I lost three months of my work. So I might have to leave the lab and the country without being able to publish some of my findings."

Experimental impacts

Eighty per cent of respondents say that the pandemic has hampered their ability to conduct experiments. One of those is Rakesh Dhama, a photonics engineer at Bangor University, UK. He was meant to travel to France earlier this year to finish experiments on a chip

THE SITUATION IS GRIM

Free-text comments in *Nature*'s survey of postdoctoral researchers worldwide detailed the downsides — and a few of the upsides — of the coronavirus pandemic. Researchers in the United States were especially vocal.

- My normal job has effectively halted during the lockdown period. However, I have been able to lead a project looking at COVID-19 as senior author, independent of my lab or institute. *Cell biologist, United Kingdom*.
- Postdoc-ing has a lot of high points… but the lack of job security, life stability and job prospects, especially with the COVID-19 recession to come, tips the overall balance far into the negative. *Food-sustainability researcher, United States*.
- I am immunosuppressed and have an autoimmune disease that puts me at high risk of COVID-19. I'll have to continue working at home for the rest of my postdoc. I'm already losing opportunities to network and prepare for the faculty job market. *Social psychologist, Canada*.
- My interactions with colleagues and professional development opportunities have been positively impacted by COVID-19. Now I have more access to opportunities that are offered online. *Ecologist, United States*.
- The pandemic has directly changed my career plan. I was going to do my postdoc in the United States, but now I am stuck in China. *Biomedical scientist, China*.
- I wish there was more intentional training from my PI about career prospects… the situation in academia may be very dire due to the COVID-19 crisis. I am not willing to wait more than two years for a faculty position. *Geneticist, United States*.
- Due to COVID, my current postdoc could end unwillingly, due to the 'hire Americans first' attitude of the US government. My visa may not be renewed, and opportunities in my home country (Canada) are non-existent, especially during COVID. *Agricultural scientist, United States*.
- Due to travel restrictions in the COVID-19 era, I am unable to join an offered postdoc position abroad. Now I am unable to get a new position. Hardly anyone wants to hire a foreign expert. *Engineer, India*.
- We are literally seen as research machines and our health and safety during the COVID-19 pandemic is left up to the discretion of our PIs. Several of my friends have been forced to work as if nothing has changed. *Quantitative health scientist, United States*.
- I love my job, I'm lucky and persevering. But COVID-19 won't allow me to have the hours of investigation I would like to have. *Marine biologist, Argentina*.
- No amount of Zoom meetings or WebEX calls can replace the feeling of going into a laboratory setting and conduct research alongside colleagues. COVID-19 did not change my research goals or career dreams, but now I feel those dreams are out of reach. *Neuroscientist, United States*.
- Under the epidemic, the employment situation is grim. *Biomedical scientist, China*.

designed to kill cancer stem cells. "Everything was scuttled because of the coronavirus," he says. "Now I won't get any credit for planning that experiment." He adds that his supervisor had acquired two pieces of equipment that could improve the accuracy of experiments, but says that no one is around to get the devices up and running. "Scientifically, coronavirus has really affected me," he laments.

Dhama, who is from India, says that his UK visa was set to expire at the end of July, adding extra urgency to a job search that was already hampered by the pandemic. With the clock ticking, he applied for a Marie Curie fellowship from the European Commission in his field of photonics. "I had to put together a 10-page proposal on a new idea in 20 days," he says. The proposal was accepted, and Dhama will start his fellowship at Tampere University in Finland in October, provided that he can get a visa to work in that country.

Experiments aren't the only scientific activities that can suffer during a pandemic. Fifty-nine per cent of respondents said that they had more trouble discussing ideas with their supervisor or colleagues, and 57% said that the pandemic had made it harder to share their research findings. A molecular biologist in the United States commented, "I haven't met my colleagues yet because of the coronavirus."

> **"I'm at an important point in my research career, and I'm not as competitive as I would have liked to have been."**

Despite the widespread delays caused by the pandemic, slightly less than 10% of respondents say that they have received an extension on their fellowships or work contracts. Nearly two-thirds (63%) say that the duration of their position has remained unchanged, and 19% were currently unsure. Melania Zauri, a cancer biologist with a Marie Curie fellowship at the Spanish National Cancer Research Centre in Madrid, says that she was given the opportunity to take unpaid leave but was not offered a paid extension of her contract. Zauri notes that Spain is extending the contracts of many researchers supported by the government, but that researchers with prestigious external fellowships are left out. "We are being treated as the last wheels on the carriage," she says.

Strained relationships

The survey included questions about supervisors, a role that takes on extra importance during a crisis. More than half (54%) of respondents said that their supervisor had provided clear guidance on managing their work during the pandemic, but one-third (32%) said that they weren't receiving that sort of support from above. Twenty-nine per cent of respondents strongly or somewhat disagreed that their adviser has done everything they can to support them during the pandemic. Female respondents (28%) were more likely than male respondents (25%) to think that their supervisors fell short.

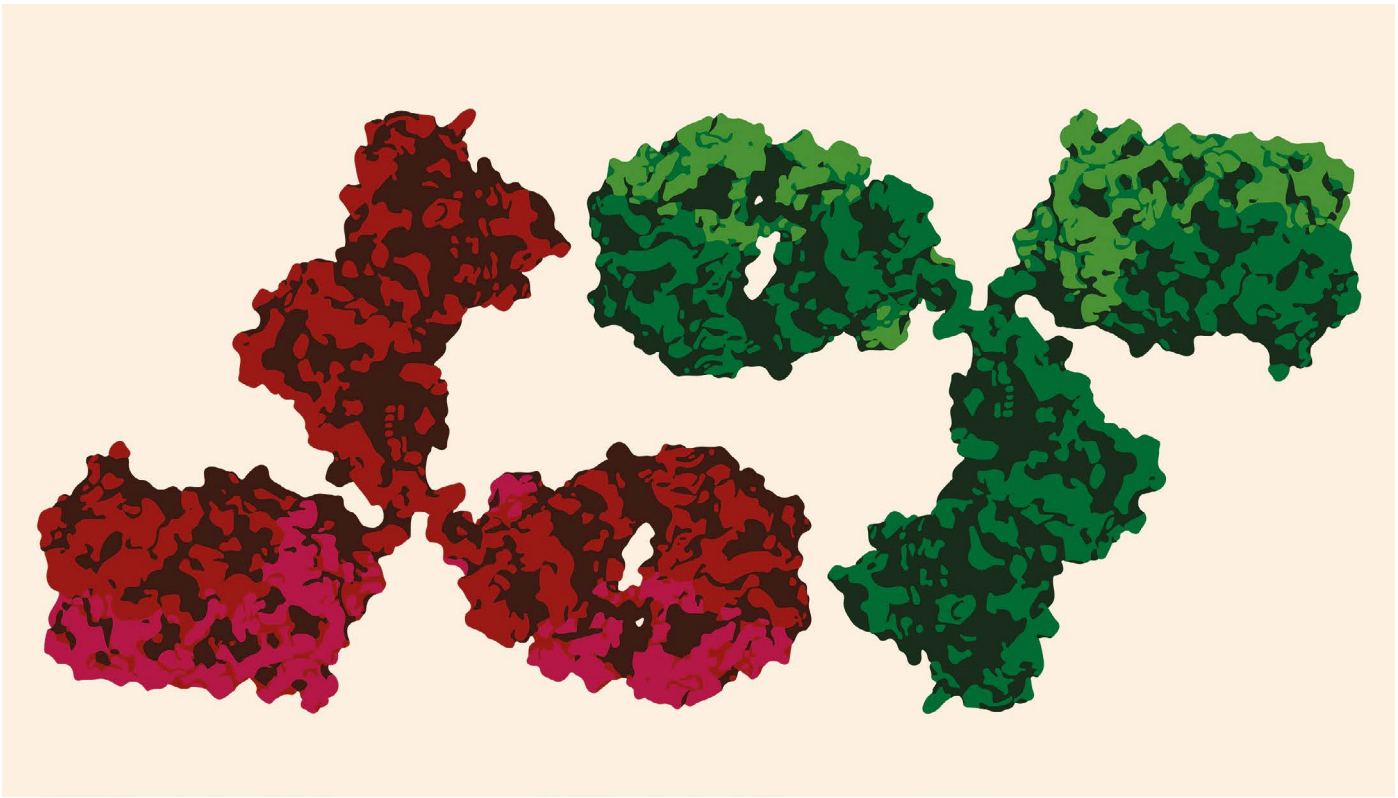
The free-comment section of the survey underscores how the pandemic has strained some supervisor–postdoc relationships. A molecular microbiologist in the United States expressed her concern about safety protocols during the outbreak. "My principal investigator pretended nothing was going on during the COVID-19 quarantine," she wrote. "He requested everybody to keep working and he refused to wear a face mask until the university made it mandatory." In a similar vein, a mycologist, also in the United States, said that lab members were "forced to continue to work with a lack of secure measures".

Some postdocs have found small consolations in the pandemic. Although more than one-quarter (26%) of respondents say that the pandemic has somewhat or significantly impaired their ability to write papers, 43% say that writing has become easier. "The downtime has allowed me to focus on my writing," Wardill says. "It's a bit of a silver lining."

Still, Wardill thinks that the pandemic has put the brakes on her work and career. As travel concerns grew during March, she felt forced to leave an ongoing research project at the University of Groningen in the Netherlands to return home to Australia. She was hoping the results and papers from that project would give her an edge as she applied for future funding, but now those experiments are on ice. "I'm at an important point in my research career, and I'm not as competitive as I would have liked to have been," she says.

Wardill hopes that funders will take the pandemic into account when assessing the research outputs and productivity of applicants. They should acknowledge the impact," she says."This is something that's affecting everyone."

Chris Woolston is a freelance writer in Billings, Montana.



ADAPTED FROM MOLEKUUL/SPL

Research antibodies are designed to recognize and bind to specific proteins according to their shape and chemical properties.

WHEN ANTIBODIES MISLEAD: THE QUEST FOR VALIDATION

Research antibodies don't always do what it says on the tin. Test for true signals before you start your experiment. **By Monya Baker**

Commercial antibodies are commonplace in biology laboratories. Researchers use these giant Y-shaped proteins to detect specific molecules in cells, tissues and test tubes. But sometimes the proteins detect other molecules, too – or even instead. When that happens, confusion can snowball.

Consider the gene *CR9ORF72*. It's often mutated in people with the neurodegenerative diseases amyotrophic lateral sclerosis and familial frontotemporal dementia. But what it actually does has been hard to pin down, partly because the widely varying locations of the protein in the cell offer more confusion than clarity.

Peter McPherson, a neuroscientist at McGill University in Montreal, Canada, suspects that the multiple locations arise from what is often seen as a trivial decision for detecting the protein: the choice of antibody. Antibodies work by binding to specific parts of a protein, according to the protein's shape and chemical properties, but an antibody produced to bind to one protein can often bind to another, and sometimes with better affinity.

That's borne out in McPherson's work. He and his team bought 16 antibodies marketed to detect CR9ORF72. Then they took a cell line that produces the protein at high levels and used the genome-editing tool CRISPR–Cas9 to make a line in which *CR9ORF72* was knocked out, so the protein would not be present. They then assessed how the antibodies performed in the two lines in a series of common tests and found that the antibody that had been used in the most publications (and cited most often) found the protein even when it wasn't there. Those that worked best for each assay had not appeared in the literature at all[1].

Others have reported comparable experiences. Cecilia Williams, a cancer researcher at the KTH Royal Institute of Technology in Stockholm, tested 13 antibodies to try to untangle conflicting data about estrogen receptor β, a protein discovered in 1996 that is a potential anticancer target. Twelve of the antibodies, including the two most popular, gave either false positives or false negatives, or both, she and her team reported[2]. "Don't take either the literature or the antibody for granted," she warns.

Researchers often buy antibodies according to the number of times the product has been cited in the literature, but that strategy can overlook newer products that have been put through more rigorous tests. They also tend to assume that others who used the antibody before them checked that it worked as intended, and that it will therefore work in their own experiments, opening the door for self-perpetuating artefacts.

"When I look at papers in general, I get depressed by the quality of the antibody characterization," says Simon Goodman, a science consultant at the Antibody Society, a not-for-profit professional association. Goodman is based in Darmstadt, Germany, and has organized a series of educational webinars on appropriate techniques for the society[3]. "If you ask 'how did you validate the antibody?',

researchers will say, 'well we bought it and the producer says that it behaves like this.'"

Often, the data that companies provide to show an antibody works come from a cell line that has been engineered to express the protein at levels substantially higher than under physiological conditions. Researchers would do better to check that an antibody can detect the protein at physiological levels, in the technique and tissue type they plan to use and, ideally, that the signal fades or disappears when levels of the protein do.

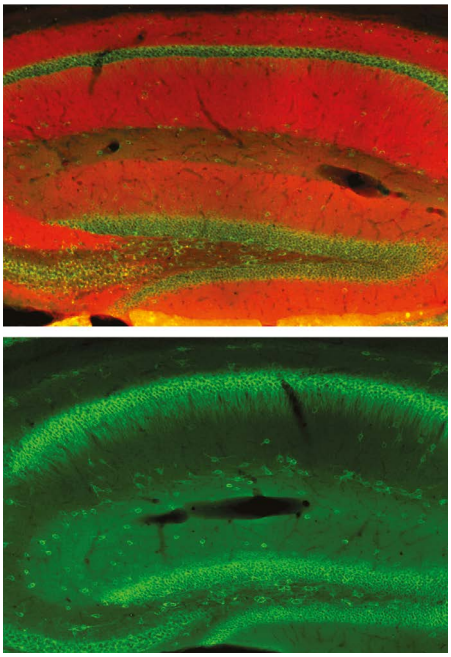
Validation drive

There has been a steady drumbeat of efforts to make researchers more careful. In 2016, the US National Institutes of Health (NIH) began requiring grant applicants to describe how they would authenticate antibodies and other key resources. Validation road maps have been printed, summits held and web portals established — Antibodypedia, Antibodies-online, Antibody Resource, Biocompare, CiteAb and Labome, to name a few.

"The true game-changer has been CRISPR," says Aled Edwards, who leads the Structural Genomics Consortium from Toronto, Canada, a public–private partnership devoted to doing basic science that can promote drug discovery. That's because the technique makes it easy to perform useful control experiments, just as McPherson and his team did. Earlier this year, the antibody vendor Abcam in Cambridge, UK, introduced a suite of knockout cell lines and preparations that researchers can buy alongside its antibodies to test positive and negative cells under specific conditions in their own labs. The company now has more than 1,600 cell lines and 2,400 cell lysates available.

Edwards and McPherson helped to set up a Toronto-based charity called YCharOS (pronounced Ikaros), to put commercial antibodies to the test. They plan to use McPherson's strategy to assess more antibodies against other targets, gauging performance across three common assays: immunoblot, immunoprecipitation and immunofluorescence. They are also working with several antibody suppliers and pharmaceutical companies to develop standard operating protocols. As well as some in-kind corporate contributions, the NIH and the Parkinson's disease charity the Michael J. Fox Foundation in New York City are providing initial funding of about US$300,000 to test a suite of antibodies used in neuroscience.

Not every antibody can be tested using knockout controls, Edwards admits. About 10% of genes are essential to life, so a knockout cell line is not viable for them. Also, an antibody that performs well in one cell line could fall short in another. Still, these simple experiments can help to identify those antibodies that aren't binding with their target protein. There are other methods researchers can try, too, such as coupling immunoprecipitation to mass spectrometry to see what proteins the antibody binds to[4]. Ultimately, says Edwards, "the onus is on the experimenter".



An antibody signal in wild-type cells (top) should disappear in tissue in which the target protein has been knocked out (bottom).

That said, it takes more than just the right antibody to yield informative experiments, says James Trimmer, who directs the NeuroMab lab at the University of California, Davis, an effort to produce high-quality antibodies for neuroscience. An antibody that works reliably when a protein is in its folded ('native') state inside a cell can perform differently when proteins are chemically altered in preserved tissue or unfolded in cell mixtures, and even small changes in sample preparation can have a large impact.

> **"The onus is on the experimenter."**

Researchers need to know how their own methods compare with those used in validation experiments, and should avoid antibodies if the validation details are unavailable. "If you use them for the wrong purpose, they won't be a good fit," Trimmer says.

It is not uncommon for labs to buy several antibodies and select the one that works best. But many developers license their antibodies to multiple distributors, who do not always disclose the antibodies' origins. When setting out to test antibodies for C9ORF72, McPherson's postdoc Carl Laflamme used CiteAb and the research literature to identify more than 100 antibody products. He then sent enquiries to vendors and scoured data sheets to rule out duplicates. Even so, the team realized later that 2 of the 16 antibodies they purchased from different companies were the same, so they had run a whole set of experiments unnecessarily. "We wasted our money and our time and effort," McPherson says.

Identity crisis

Sometimes it's not even clear which antibody researchers have used, especially in older studies. Only about 11% of the antibodies used in papers published in 1997 are identifiable, according to an analysis[5] led by researchers at the University of California, San Diego (UCSD), and the data-sharing platform SciCrunch in La Jolla, California; nowadays, that figure has risen to 43%.

Anita Bandrowski, head of SciCrunch and a bioinformatician at UCSD, is spearheading an effort to assign every antibody a unique identifier, called an RRID, and include it in publications. Researchers can find or request an RRID on the Antibody Registry's website, and the identifiers would remain the same even if the vendor supplying an antibody changes its catalogue or goes out of business. Antibodies are much easier to find when journals mandate RRIDs, says Bandrowski. The journal *Cell*, for instance, asks authors for RRIDs, and 97% of its antibodies are findable[5]. Both *Nature* and Nature Research journals encourage the use of RRIDs to track key biological resources, including antibodies, cell lines, model organisms and tools.

RRIDs can alleviate, but not solve, the problem of the same antibody being sold by many vendors: if the original source is clearly disclosed, all the antibodies can be assigned the same RRID. Bandrowski guesstimates that the 2.5 million antibodies with RRIDs represent perhaps 700,000 unique molecules. But RRIDs do not distinguish between different batches of the same product, which can be particularly problematic for polyclonal antibodies, which are purified from the blood of immunized animals and are therefore more of a mixture than those made from cultured cells.

The bottom line is: however an antibody-driven experiment comes out, researchers would be wise to be sceptical. When experiments fail, researchers often question their own technique, says Goodman. "Of course you blame yourself as a young scientist." But the scientific community should be equally sceptical of antibodies that seem to work, says Edwards, and demand evidence that they do before relying on them. "We buy antibodies, we don't test them, and then we publish articles that send the field sideways."

Monya Baker is a senior Comment editor at *Nature*.

1. Laflamme, C. *et al. eLife* **8**, e48363 (2019).
2. Andersson, S. *et al. Nature Commun.* **8**, 15840 (2017).
3. Voskuil, J. L. A. *et al. mAbs* **12**, 1794421 (2020).
4. Uhlen, M. *et al. Nature Meth.* **13**, 823–827 (2016).
5. Menke, J., Roelandse, M., Ozyurt, B., Martone, M. & Bandrowski, A. Preprint at bioRxiv https://doi.org/10.1101/2020.01.15.908111 (2020).

Where I work
Peter Tsai

Photograph by Carlos Jones/ Oak Ridge National Laboratory.

I spent my career working with microfibres, which have a diameter of 2 micrometres or so. Tangled mats of microfibres made of polypropylene are good materials for air filters — such as those in N95 respirators. They filter at least 95% of airborne particles and meet US National Institute for Occupational Safety and Health standards for medical use. In 1992, I invented a way to charge those filters by passing them through a device that produces static electricity — much like when you rub a balloon on your hair, but it's permanent. The electrostatic charge makes the filters ten times better than uncharged filters at blocking viruses and other particles.

In this picture, I'm at the Oak Ridge National Laboratory in Tennessee, where I recently helped convert a system for making a precursor to carbon fibres into one that can produce and charge polypropylene filters for N95 masks. We melt the polypropylene and it's extruded, like spaghetti. The extruder produces hundreds of fibres at once, then blows hot air at them so they tangle into a web as they land on the passing conveyer belt. Then they pass through the charging device — I gave them one I had.

The laboratory, which can now produce material for 9,000 masks an hour, aims to share the technology with other labs for research, and to train companies to make the material.

I retired in 2018, but have stayed busy, especially since the COVID-19 pandemic began and face masks became so important. People began to contact me because of the N95 shortage. They wanted to sterilize the masks for reuse without damaging them or destroying the electrostatic charge. I knew that heat would not alter the charge, but that alcohol would erase it. I experimented and learnt that ozone sterilization would retain the charge, but that ozone would crack natural rubber straps.

Even though I'm mostly working for free, I feel a responsibility to help out during the pandemic. Otherwise, I would regret it for the rest of my life.

Peter Tsai is a materials scientist in Knoxville, Tennessee. **Interview by Amber Dance.**